

# A

## Abiotic Stress Tolerant Crops: Genes, Pathways and Bottlenecks

KORNEEL VANDENBROUCKE, MICHAEL METZLAFF  
Bayer BioScience N.V, Ghent, Belgium

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Abiotic Stress-Tolerant Crops  
Limitations of Genetic Engineering for Stress Tolerance  
Systematic Phenotyping of Plants  
Future Directions  
Bibliography

### Glossary

**Abiotic stress** Any negative impact on a living organism caused by a nonliving factor in the environment in which the organism is present.

**Biological modeling** *In silico* description of a biological process to generate predictions for experimental validation.

**Conventional breeding** Development of new plant varieties by selection after natural reproduction.

**Epigenetics** The study of inherited changes in phenotype or gene expression caused by mechanisms other than changes in the DNA sequence.

**Genetic engineering** Development of a new plant variety through genetic modification by using recombinant DNA technology.

**Phenotype** Observable trait or characteristic (i.e., appearance) of an organism in a specific environment.

**Phenome** Collection of phenotypes of an organism in all possible environments.

**Phenotyping** Process of studying the phenotype.

**Plant productivity** Ability of a plant to produce a certain amount of biomass, either as green tissue and/or as seeds (yield).

**Quantitative trait locus (QTL)** Stretches of DNA that are closely linked to the genes that underlie the trait in question.

**Regulon biotechnology** Genetic engineering by targeting genes encoding proteins with regulatory function, often transcription factors.

**Trait** Characteristic of an object.

### Definition of the Subject

World food and feed security is increasingly dependent on continuous crop improvement and, in particular, the development of crops with increased resistance to abiotic stresses. This economical and social challenge has attracted the global community of plant breeders and scientists and many potential solutions have been put forward. Our understanding of the response of plants to abiotic stress has significantly improved over the last year. However, abiotic stress tolerance is a complex trait that can be affected by many external factors. Abiotic stress tolerance involves many processes that are not yet completely understood and several limitations still need to be overcome. Recent advances in many areas of plant research, including phenotyping, make scientists optimistic that valuable solutions will be found to allow deployment/commercialization of plants better able to tolerate abiotic stresses.

### Introduction

A growing world population with increased living standards combined with the urgent need for a more sustainable agriculture demands the development of crop varieties which are able to cope with fluctuating and adverse environmental conditions limiting plant growth and productivity, referred to as abiotic stresses [1]. Further global warming is expected to aggravate the negative impact of abiotic stresses. Low water

availability (drought stress), high salinity (salt stress), and high temperatures (heat stress) are considered to be among the most threatening abiotic stresses, which can compromise up to 80% of the attainable yield [2]. New technologies for plant improvement that aim to overcome the negative impact of these stresses therefore need to be urgently identified and implemented. Abiotic stresses and their effects on plant productivity have attracted academic scientists as well as large and small Ag biotech companies. This is reflected by increasingly more publications in peer-reviewed journals (including recent issues in plant physiology and functional plant biology dedicated to abiotic stress), online centralization of relevant information ([www.plantstress.com](http://www.plantstress.com); [www.yieldbooster.org](http://www.yieldbooster.org)) and numerous patent applications protecting potential new solutions (Fig. 1).

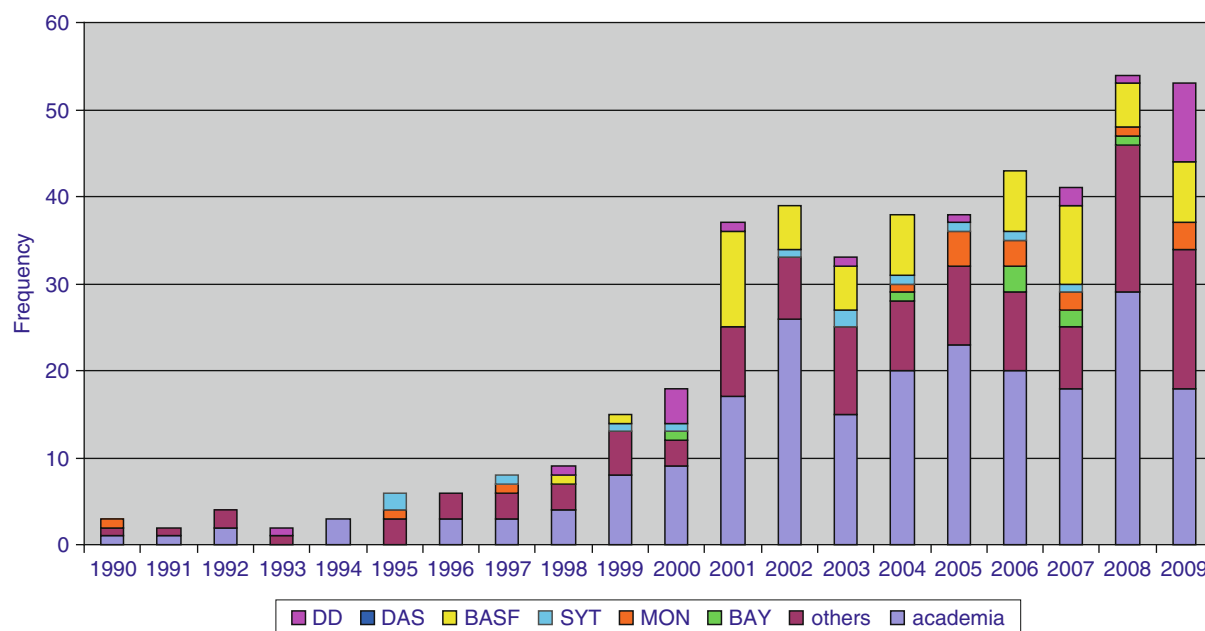
Plants respond to abiotic stresses at different levels in an avoidance-tolerance mode, which includes physiological and molecular changes (Fig. 2). An important interest of plant biologists aims not only to better understand the adaptive response of plants to abiotic stress, but also to exploit this knowledge for the production of varieties that are better protected against

the detrimental effects. Before the genomics era, classical or conventional breeding of plants with improved physiological characteristics was the only way to improve crop productivity. In the last 20 years, mapping of quantitative trait loci (QTL), genetic engineering, and their implementation in marker-assistant (molecular) breeding have become increasingly important [3–5]. Such genomics-based approaches rely on the identification of genes (gene discovery) that may be valuable candidates for crop improvement and were empowered by the advent of state-of-the-art molecular tools, such as DNA sequencing and expression profiling [6].

In this entry, some of the recent genomics-based advances in engineering stress-tolerant crops are summarized, existing limitations associated with these approaches are described, and some emerging trends facilitating better evaluation of crop performance in the greenhouse and in the field are outlined.

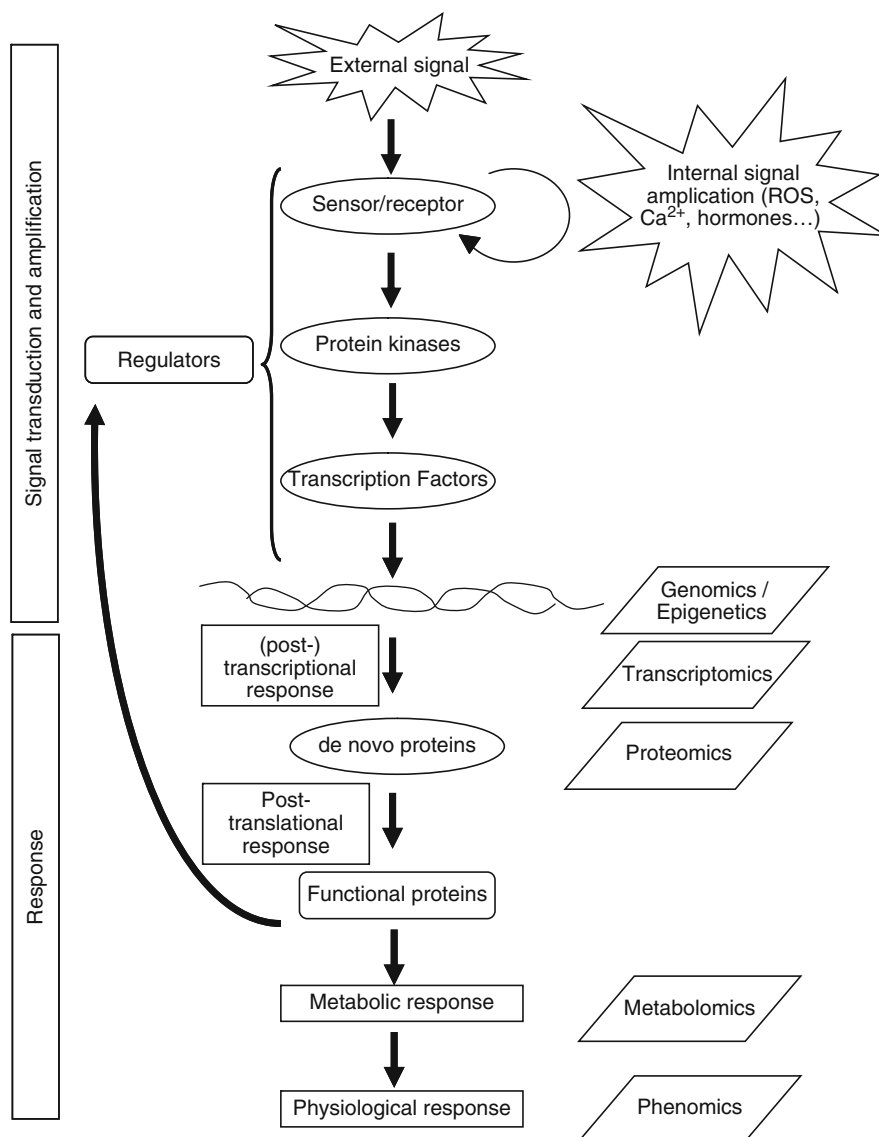
### Abiotic Stress-Tolerant Crops

Plants protect themselves from the detrimental effects of abiotic stresses by increasing the expression of



**Abiotic Stress Tolerant Crops: Genes, Pathways and Bottlenecks. Figure 1**

Abiotic stress-related publications of international patent applications by academia and agbiotech companies (provided by Rolf Deblaere). BAY Bayer, MON Monsanto, SYT Syngenta, DD Dupont, DAS Dow AgroSciences



**Abiotic Stress Tolerant Crops: Genes, Pathways and Bottlenecks. Figure 2**

Signal transduction and response of plants to abiotic stress with the corresponding evaluation platforms

defense or stress tolerance genes. Plant varieties that evolved to have constitutive or high(er) expressed levels of such genes are better adapted to abiotic stress conditions. Genes with potential value for genetic engineering of stress tolerance have been ever more discovered either indirectly by genetic dissection of identified QTLs in stress-tolerant varieties, or directly through their changed expression in plants exposed to stress. Introduction of selected genes seems to be the favored way to improve plants in the future [1]. Plant

biotechnologists have been reporting genetically engineered plants with increased stress tolerance for almost 2 decades. Many genes have already been directly employed to improve stress tolerance of higher plants using a wide variety of approaches [7].

One attractive approach for engineering stress tolerance in crops has been the constitutive overexpression of stress genes from bacteria, yeast, or model plant species, such as *Arabidopsis thaliana* (*Arabidopsis*). Such a heterologous gene approach

worked particularly well when the genes encode proteins with single biological functions that are often absent or only expressed at low levels in the target plant. Such genes are often referred to as functional genes and can encode molecular protectants, detoxifying proteins, or ion transporters. A second group is comprised of regulatory proteins with often multiple biological functions, including enzymes involved in (phospho)lipid signaling, protein kinases, protein phosphatases, calcium/calmodulin-binding proteins, and various transcription factors (TFs) [8]. With the emergence of large-scale genome sequencing in crops, several crop orthologues of stress genes could be identified and increasingly be used to engineer stress tolerance in crops. Here some of the approaches for engineering stress tolerance using genes involved in different cellular processes or pathways are summarized, including detoxification, protein stabilization, osmoregulation, transport, lipid metabolism, transcription and signaling, and posttranscriptional, and (post-) translational regulation. However, genetic engineering for stress tolerance in crops is not limited to these gene classes as recently shown for mustard annexin Bj1, a gene encoding a calcium-dependent phospholipid and cytoskeleton binding protein that is involved in golgi-mediated secretion, which resulted in stress tolerance when expressed in cotton [9].

### Detoxifying Genes

Most, if not all, abiotic stresses induce the accumulation of reactive oxygen species (ROS), which in turn cause oxidative stress [10, 11]. ROS are extremely reactive, and therefore can undergo uncontrollable and damaging reactions with cellular components, including DNA, lipids, and proteins. This can aggravate the detrimental effects of the initial stress and can even lead to cell death [12, 13]. To protect against oxidative stress, plant cells possess an extensive ROS scavenging network, which involves nonenzymatic antioxidants, including vitamin C, vitamin E, glutathione, carotenoids, and flavonoids, as well as numerous enzymatic mechanisms such as multiple superoxide dismutases (SOD), catalases (CAT), ascorbate peroxidases (APX), glutathione peroxidases (GPX), glutathione-S-transferases (GST), alternative oxidases, and peroxiredoxines [12]. It was hypothesized that alleviation of oxidative damage by

the use of ROS scavengers would enhance plant resistance. This was confirmed by a number of transgene transfers using this detoxification strategy. Stress tolerance could be improved by either direct scavenging of ROS or by enhanced removal of oxidative damaged and hazardous components accumulating in the cell. Since the accumulation of ROS and derivatives thereof is a common theme during most, if not all, abiotic stresses, the detoxification strategy enabled the generation of transgenic plants with simultaneous tolerance to multiple stresses [14–22].

In crops, enhanced stress tolerance was achieved by increasing the level of typical scavenging enzymes such as GST, different SOD isoforms, APX, and CAT [23–28]. Overexpression of a GST enzyme in rice resulted in increased protection against salt, low temperature, and oxidative stress [23]. SOD catalyzes the dismutation of superoxide into oxygen and hydrogen peroxide ( $H_2O_2$ ). Transgenic rice plants expressing a manganese SOD from pea were more tolerant to drought stress [26]. Similarly, transgenic oilseed plants ectopically expressing a wheat manganese SOD were tolerant to oxidative and heavy metal stress [15]. A copper/zinc SOD was shown to be effective to protect rice plants against drought, salt, and oxidative stress [28]. Excess  $H_2O_2$ , produced either as by-product of superoxide detoxification by SOD or directly by metabolic processes, can be removed by APX, which catalyzes the reduction of  $H_2O_2$  to water. Cotton plants with high levels of an APX isolated from pea were tolerant to low temperature stress [24]. Because of their complementary functions (i.e., removal of superoxide and  $H_2O_2$ ), also the combination of SOD and APX resulted in enhanced stress tolerance [27]. As APX, also CAT protects plants from toxic  $H_2O_2$  molecules. Hence, expression of a wheat CAT enzyme in rice protected these plants against low temperature stress by reducing the levels of  $H_2O_2$  [25].

Engineering tolerance to oxidative stress is not limited to the use of the traditional ROS detoxifying enzymes. Significant improvement of stress tolerance in tobacco plants was achieved by overexpressing a stress-responsive aldehyde dehydrogenase gene from maize [29]. Recently, it was also shown that ectopic expression of the cotton stress-responsive MT3a gene in tobacco increased tolerance to salt, drought, and low temperature [30]. MT3a belongs to the



metallothionein family that has numerous cellular functions including the regulation of metal homeostasis and oxidative stress. Another gene that can protect plants from abiotic stress-induced ROS is the mitochondrial alternative oxidase (AOX), the terminal oxidase in the alternative respiratory pathway of plants [31]. For example, tobacco AOX1a is necessary for survival against oxidative stress when the cytochrome pathway is dysfunctional [32].

### Protein Stabilization by Molecular Chaperones

One major detrimental effect of abiotic stresses is that they usually cause protein dysfunction through denaturation and aggregation of nonfunctional or aberrant proteins. Maintaining proteins in their functional conformation is therefore important for cell survival under stress. This can be accomplished, for example, through transcriptional induction of genes encoding heat shock proteins (HSPs, [33]). HSPs control the correct folding and conformation of both structural (e.g., cell membrane) and functional (e.g., enzymes) proteins. This important function has prompted researches to create transgenic lines with increased HSP levels. Studies on HSP proteins in plants have mostly focused on heat stress [34–42]. Although most of the evidence is limited to *Arabidopsis*, the protective capacities of HSP proteins and their potential economical value for crop engineering for heat stress tolerance was proven in tomato [42]. Recently, it was demonstrated that constitutive expression of a cotton HSP, GHSP26, enhanced drought tolerance in transgenic cotton plants [43].

In addition to HSPs, also LEA-type proteins can confer molecular protection of cellular components during abiotic stress [44]. LEA-type proteins are encoded by RD (responsive to dehydration), ERD (early responsive to dehydration), KIN (cold inducible), COR (cold regulated), and RAB (responsive to abscisic acid) genes in different plant species [45, 46]. As HSPs are typically induced by high temperatures, LEA proteins accumulate in response to dehydration (drought, osmotic, and/or cold stress). The actual functions of these proteins remain however largely unknown. Their hydrophilic nature suggest that LEA proteins act as water-binding proteins, but additional functions, including ion sequestration and protein and membrane stability, have also been proposed [47, 48].

Few examples for the use of LEA proteins to engineer stress tolerance in crop exist. Increasing the levels of endogenous LEA3 through genetic engineering of rice made these plants more tolerant to drought stress [47]. Ectopic expression of barley HVA1 in oat resulted in enhanced salt and osmotic stress tolerance [49].

### Osmoregulation and Protection by Genes Involved in Metabolite Biosynthesis

Plants respond to drought stress by producing organic compounds to avoid water loss from cells (dehydration) and damage to essential components (osmotic stress). Therefore, one of the earliest approaches for genetic engineering of stress tolerance in plants (reports dating from the early 1990s) consists in enhanced synthesis of such metabolites, called osmoprotectants [50, 51]. Osmoprotectants include sugars and sugar alcohols (e.g., mannitol, trehalose, and galactinol), amines (e.g., polyamines and glycine betaine), and amino acids (e.g., proline) [52, 53]. These molecules normally do not interfere with cellular functions and are therefore often referred to as compatible solutes. Many plant species lack the ability to synthesize the special osmoprotectants that naturally accumulate in stress-tolerant species. Therefore, several transgenic approaches to increase the synthesis of osmoprotectants used bacterial biosynthetic genes, such as CodA and BetA (glycine betaine), MtlD (mannitol), and genes from the ectoine or trehalose biosynthesis operon [54–59]. Alternatively, key biosynthetic genes, including betaine aldehyde dehydrogenase and choline monooxygenase (glycine betaine biosynthesis), and pyrroline carboxylate synthase (proline synthesis), were isolated from specific plant species, such as *Vigna aconitifolia* and *Spinacia oleracea* and used to induce drought and salt tolerance in wheat and rice respectively [60, 61]. Although the accumulation of compatible solutes during stress is mainly important for osmoregulation and for maintaining correct protein structures, this may also be important for reducing or preventing the damaging effects of ROS [62, 63].

### Transport Proteins

Ion transport proteins are involved in reestablishing ionic homeostasis after salt stress either by increasing ion storage in the vacuole, or by improving ion

excretion from the cells [64]. Different types of ion transporters, depending on their localization and selectivity, have been the target of genetic engineering. These include both vacuolar and membrane  $\text{Na}^+/\text{H}^+$  antiporters, vacuolar  $\text{Ca}^{2+}/\text{H}^+$  antiporter, and  $\text{Mg}^{2+}$ ,  $\text{Na}^+/\text{K}^+$ , and  $\text{Ca}^{2+}$  transporters [43, 64]. Known salt stress tolerance genes encoding ion transporters isolated from crops include NHX1 (rice), KAT1 (rice), and HKT1 (wheat) [65–67]. Plasma membrane cation/proton antiporters cause alkalinization of the apoplast, thereby changing the activity and conformation of membrane proteins, which might serve as a signal to mediate gene regulation and induce a general stress response [68]. Besides using ion transporters to induce tolerance to salt stress, tolerance to osmotic stresses was engineered by increasing the levels of proteins involved in water transport [69]. Tolerance to heavy metal stress was achieved by constitutive expression of heavy metal transport proteins, including wheat ALMT1 [70–72].

### Lipid Metabolism

Adaptation of living cells to low temperatures involves alterations in the membrane lipid composition, for example, by decreasing membrane fluidity through fatty acid unsaturation. It was demonstrated that increasing the number of unsaturated fatty acids by genetic engineering could improve stress tolerance in plants [73–77]. Overexpression of a spinach glycerol-3-phosphate acyltransferase (GPAT) in rice increased tolerance to low temperature whereas downregulation of a fatty acid desaturase in rice increased tolerance to high temperatures [73, 78]. Upon overexpression, tomato GPAT also increased the tolerance of tomato plants to low temperature stress [76].

### Stress Sensing and Signal Transduction

The first and probably most important step in the response of plants to abiotic stress is the sensing or detection of the external stimuli by receptors, typically located in the cellular membranes. The identification of important stress receptors has been a difficult challenge and until recently, there was no report on their use in engineering stress tolerance. Overexpression of *Arabidopsis* membrane-bound receptor-like protein kinase 1 and a membrane located calcium/calmodulin-binding receptor-like kinase from soybean in

transgenic *Arabidopsis* resulted in enhanced plant tolerance to drought and salt stress, respectively [79, 80]. In rice, overexpression of SIK1 resulted in higher tolerance to salt and drought stresses [81]. Other receptors operate in the cytosol and react to internal stimuli that amplify abiotic stress signals, such as calcium and ROS, as well as various hormones and other small molecules. The best-studied stress hormone is abscisic acid (ABA), which is involved in stomatal closure and plays a crucial role in tolerance against drought stress by preventing transpiration and water loss from stomata [82]. ABA also has an essential role in activating signal transduction pathways involved in tolerance to drought, cold, and salt stress. Recently, an important ABA receptor protein family has been identified [83, 84].

Stress receptors or sensors transduce external and internal signals into an intracellular response, for example, through phosphorylation and dephosphorylation cascades controlled by protein kinases and phosphatases. Genes encoding protein kinases were successfully exploited to engineer stress-tolerant crops. Modification of normal endogenous levels of GSK1, SAPK4, CDPK7, CIPK03, CIPK12, and CIPK15 in rice protected the plants against various abiotic stresses, including drought, salt, and low temperature stress [85–88]. Expression of tobacco mitogen-activated protein kinase, NPK1, in maize resulted in drought and low temperature stress tolerance [89]. Constitutive overexpression of a stress-inducible small GTP-binding protein PgRAB7 from *Pennisetum glaucum* enhances abiotic stress tolerance in transgenic tobacco [90].

### Transcriptional Regulation

In the last decade, the most widely used and probably most important strategy for engineering abiotic stress tolerance in plants relied on the expression of genes that are involved in signaling and regulatory pathways [91, 92]. The use of TFs for tailoring stress tolerance is often referred to as regulon biotechnology because it affects the expression of many target genes in parallel [93, 94]. One of the reasons for their popularity is that TFs are believed to mediate durable tolerance to multiple stresses. Most TFs that control stress tolerance in plants belong to (large) protein families based on the

presence of common DNA binding motifs and selectivity toward certain cis-regulatory elements in the promoters of target genes. These families include AP2/EREBP/ ethylene responsive element binding proteins (AP2/EREBP) such as ethylene responsive factors (ERF), the DREB/CBF (drought-responsive element binding/cold-responsive element binding factor) proteins, basic domain leucine-zipper (bZIP) proteins such as ABFs (abscisic acid (ABA)-responsive element binding factor), basic helix-loop-helix proteins (including MYC proteins), NAC (petunia NAM *Arabidopsis* ATAF1/2, and CUC2-domain) proteins, MYB-related proteins, as well as different families of zinc-fingers domain-containing proteins, such as WRKY binding factors, C<sub>3</sub>H- and C<sub>2</sub>H<sub>2</sub>-type TFs. In recent years, many excellent reviews have been published on the role and use of TFs for engineering of stress tolerance in plants [48, 93–96].

Many AP2/EREBP-type TFs, the best studied being DREB/CBF proteins, have been used to engineer stress-tolerant crops. Transgenic rice plants with higher levels of CBF1/DREB1B and CBF3/DREB1A were more resistant to drought, salt, and low temperature stress [97]. A similar approach with CBF15 and CBF17 in oilseed and NF-YB2 in maize resulted in increased tolerance to low temperature and drought stress, respectively [98, 99]. Potato plants with increased levels of endogenous EREBP1 were more tolerant to salt stress and low temperatures [100]. Ectopic expression of the soybean ERF3 gene in transgenic tobacco plants gave tolerance to drought and salt stress [101]. Similarly, ectopic expression of an *Arabidopsis* AP2/ERF TF, HARDY, in rice also induced tolerance to drought and salt [102]. Expression of TERF1 in rice regulates expression of stress-responsive genes and enhances tolerance to drought and high salinity [103].

However, not only the AP2/EREBP-type TFs have been exploited to engineer stress-tolerant crops. For example, in rice, it was shown that constitutive expression of proteins from various other TF types, including bZIP23, zinc-finger protein 245, TIFY11, MYB3R-2, IRO2, NAC6, SNAC1, and PF1 TFs, could induce tolerance to drought, salt, low temperature, and nutrient deficiency [104–111], and the rice *dst* mutant, DST encoding a novel C<sub>2</sub>H<sub>2</sub>-type TF, showed increased tolerance to drought and salt [112]. In tomato, ectopic expression of rice MYB4 and pepper PIF1 induced low

temperature and drought stress tolerance, respectively [113, 114]. Also in tomato, constitutive expression of SLAREB, a bZIP TF with affinity for ABA-responsive elements, increased tolerance to drought and salt stress [115]. Ectopic expression of cotton ZFP1, encoding a CCCH-type zinc-finger protein, and rice ZFP177, an A20/AN1-type zinc finger, enhanced stress tolerance in tobacco [116, 117].

### Posttranscriptional and (Post-)Translational Regulation

Posttranscriptional control of stress gene expression is mediated by proteins that are involved in splicing, export, and degradation of gene transcripts, which contributes to correct function of the encoded proteins. In the last years, it has become evident that nonprotein coding RNA molecules, including microRNA and other small RNA molecules, play a very important role in posttranscriptional regulation of plant stress responses [118]. miRNA-mediated posttranscriptional control of antioxidant gene expression seems very important in plants, as shown for APX during programmed cell death and drought stress, and for Cu/Zn SOD during tolerance against oxidative stress [21, 119, 120]. ROS-induced stabilization of SOS1 mRNA transcripts is essential for SOS1-dependent salt tolerance [68].

Several genes that encode proteins involved in RNA processing were discovered to be involved in stress tolerance processes [121–125]. In addition, also proteins that control translation (deoxyhypusine synthase, DHS), posttranslational modification (peptide methionine sulfoxide reductase, PMSR4), and protein degradation (SDIR1) are interesting candidates for engineering stress tolerance in plants [126–128]. However, the above examples resulted only from *Arabidopsis* research and much less information on the importance of such processes is available for crops. Recently, it was shown that transgenic rice plants with increased levels of methionine sulfoxide reductases, MSRA and MSRB, are more tolerant to salt stress [129]. Two other approaches were reported to work in both *Arabidopsis* and oilseeds. Wang and coworkers (2005) reported that loss-of-function of farnesyl transferase (FTA and FTB) increased tolerance to drought stress [130]. Similarly, reduction of poly-ADP-ribose

polymerase (PARP) activity improved drought tolerance by increasing energy use efficiency [131]. PARP is involved in the modification of nuclear genes, such as histones. Downregulation of PARP resulted in the deregulation of the expression of genes in response to stress [132]. Processes that control stress tolerance and general fitness of plants in the field, such as energy use efficiency, can even be controlled by an epigenetic component [133]. By starting with an isogenic canola line, Hauben and coworkers could separate in subsequently selfed generations “good” and “bad” performing plants, which only differed in epigenetic parameters but remained genetically identical.

All these examples show that the plant response to abiotic stress is not only dependent on transcriptional changes, but also on RNA processing, translation, post-translational processes, and epigenetics, which are nearly unexploited until today. The lack of understanding these processes is a limiting factor in engineering stress tolerance in crops.

### Limitations of Genetic Engineering for Stress Tolerance

Because conventional breeding is a very time-consuming process, genetic engineering is now an important technology for many commercial applications aiming at plant improvement [1]. Today this mainly includes the production of plants with engineered herbicide or insect tolerance. However, it has been and still is extremely challenging to progress engineered plants with reported abiotic stress tolerance from research to field applications for farmers [134]. The main factors hampering the production of commercial solutions for the abiotic stress-related problems in agriculture are related to: (1) the availability of sequence information and transformation protocols, (2) the genetic diversity between model species and crop, (3) the multigenic character of stress tolerance, (4) the definition of stress tolerance and (5) the methods for the evaluation of stress tolerance [53, 135].

The first requirements for genetic engineering of plants include the availability of sequence information and transformation protocols. This important information is still lacking for many agricultural crops. For long time, this has limited (academic) research to model plant species. Although the use of model species

such as *Arabidopsis* has proven its value, the genetic diversity within higher plants is the main factor that limits translatability of results obtained in model species into commercial crops. Even in those cases where the research is conducted in a more closely related model crop, it remains to be a challenge to translate the results into genetically diverse elite varieties used for commercialization.

Another factor limiting translatability from model species to crops is the complex, multigenic character of stress tolerance mechanisms. Despite the use of single genes for engineering stress tolerance, efficient and sustainable stress tolerance requires the cooperative action of many genes that are involved in various cellular processes that are not completely understood. Due to the initial success of transcriptomics, it has been accepted for long time that responses to abiotic stresses were almost exclusively regulated at the transcriptional level by a small set of core transcription factors. However, it is now clear that many other regulatory processes such as posttranscriptional control, translation, posttranslational, and epigenetic effects play an important role. In addition, some of these processes might not always be evolutionarily conserved but rather genetically diverse.

The early availability of the full genome sequence and many molecular analysis tools has turned *Arabidopsis* into the primary model species of choice for academic research in general, and for studying the response of plants to abiotic stresses in particular. However, the methods for evaluation of stress tolerance using *Arabidopsis* plants grown in a laboratory environment and agricultural crops in the field can be quite diverse. From an agronomic point of view, it is more relevant to study the effects of abiotic stresses on plant growth and yield over longer periods which cover the life span of a crop under field conditions. Appropriate field tests are therefore highly important because they allow studying plants in their natural environment in which they are exposed to combination of multiple stresses or cycles of the same stress [136]. In contrast to the field situation for crops, most studies on abiotic stress in *Arabidopsis* focus on short-term and strong stress treatments. Therefore, evaluation of stress tolerance is often focused on survival which is easier to screen. In contrast, yield-related traits are more difficult to measure but much more important for

crops. Abiotic stress tolerance of crops should therefore be defined as the potential to produce high yields when exposed to mild(er), multiple stresses instead of the ability to survive a single, lethal stress.

Another limiting factor for successful field applications of plants with engineered protection against stress is the seemingly mutual exclusive characteristic of high yield under normal conditions and high stress tolerance. Most reported experiments focused either on abiotic stress tolerance or on yield rather than including detailed analysis of both traits. More in-depth analysis of stress-tolerant plants under “normal” conditions often revealed a negative effect on growth, development, and yield traits. On the other hand, engineered plant with improved growth features, for example, by boosting a metabolic process, can be less adapted to maintain growth under abiotic stress conditions. Negative effects, often referred to as yield penalty, can be (at least partially) circumvented by temporally and spatially controlled expression of the target gene [137]. However, the availability of suitable promoters is still limited. Improved sequencing and bioinformatic tools for identification and construction of promoters allowing the expression of genes in a spatially and temporally controlled manner can overcome this bottleneck.

Phenotyping experiments with a set of genetically defined plants are often difficult to repeat because plants are largely influenced by the environment in which they grow. In fact, growth and yield (i.e., performance) of a plant under normal or stress conditions is the output of all integrated physiological processes in the plant. It is therefore not surprising that different laboratories often obtain different results for plant performance [138]. More effort is needed to describe and standardize the design of the experiments, of protocols for measuring traits, and of the growing conditions. Different methods to evaluate stress tolerance in *Arabidopsis* and a conceptual framework for phenotyping during breeding for drought tolerance have already been proposed, including defining the minimal information needed for carrying out a drought stress experiment [139, 140]. Improved and systematic phenotyping of plants will greatly contribute to overcome the limitations associated with the methods to evaluate stress tolerance.

## Systematic Phenotyping of Plants

In the last decades, it has become clear that there is a substantial need for automation of research processes [141]. Automation not only allows increasing throughput, but also ameliorates standardization, reproducibility, and therefore the overall quality of an experiment. Automation increases the research value and also reduces “cost to practice” for companies aiming to apply the research findings toward the development of a superior and sustainable product. It can therefore be assumed that automation will help to improve the translation of research findings into valuable products for the customer.

Plant research has already embraced the development of high-throughput screening experiments that were used for drug or herbicide discovery. Such experiments focused on assessment of certain molecular function. Next to these screens, also high-content screening protocols were developed to assess such molecules in a biological assay. Although these technologies have proven value for single trait discovery, the efficient development of plants with complex, multifactorial traits such as increased tolerance to abiotic stresses is fully dependent on protocols and methods to automatically phenotype at the plant level. The performance of whole plants is assessed by studying their physiology or phenotype in tissue culture, environmentally controlled growth chambers, greenhouse, or the field. Various methods for automated phenotyping of plants or plant parts in well-defined environments have been published [141]. After the establishment of genomics, transcriptomics, proteomics, and metabolomics in plant sciences, which all greatly contributed to gene or lead discovery in the field of abiotic stress, plant phenomics needs to be developed and implemented [142–144].

Plant phenomics is expected to enable the efficient and reliable evaluation of a new trait solution and thereby enhance breeding for stress-tolerant varieties. Large, dedicated plant phenotyping centers such as the ACPFG in Adelaide (Plant Accelerator™; <http://www.plantaccelerator.org.au/>), the CSIRO Plant Industry in Canberra (High Resolution Plant Phenomics Centre; <http://www.plantphenomics.org.au/HRPPC>), or the Research Centre in Jülich (Jülich Phenomics Centre; <http://www.fz-juelich.de/icg/icg-3/jppc/>) have been



created. These research institutions can act as service providers for academia and industry by centralizing both high-tech infrastructure and highly skilled researchers with different background, including engineering, mathematics, computer science, and plant physiology/biology. An important biological question that probably can be answered by automated phenomics studies is which phenotypes or combination of multiple phenotypes need or do not need to be collected in order to be able to correctly evaluate plant performance in the environment of interest. Moreover, parallel phenotyping of plants in one or multiple environments allows detection of associations between traits [145]. Identification of redundant information from (highly) associated phenotypes will help to reduce complexity and focus on essential phenotypes.

Well-described examples of automated platforms from academia or private companies to study the phenotype of model plants or crops include phenopsis, phenodyn, GROWSCREEN, TraitMill™, PlaRoM, and various LemnaTec products. Phenopsis is a phenotyping platform for *Arabidopsis* that allows measuring multiple parameters, including plant growth, water use, and transpiration rates, which are associated to plant performance under normal and drought stress conditions [146]. Phenodyn was specifically constructed to perform similar experiments with monocots such as maize and rice [147]. GROWSCREEN is an automated method for growth analysis of a limited set of small seedlings [148]. TraitMill™, one of the first automated phenotyping systems, was developed to study traits in rice [149]. LemnaTec offers commercialized phenotyping systems for different purposes and has customers from both academia and industry all over the world (<http://www.lemnatec.com>). Although most phenotyping platforms focus on the shoot of the plant, several systems for root phenotyping were also developed [150, 151]. By using large field scanners, it is even possible to use similar approaches during field trials.

One common feature of all automated phenotyping systems is the use of digital cameras for imaging. The use of digital imaging has many advantages over traditional plant phenotyping efforts. While traditional experiments were often dependent on subjective visual scoring of plant traits or on labor-intensive, often destructive manual sampling, for example, for

fresh and dry weight measurements, automated phenotyping by digital imaging is nondestructive, and therefore allows following the development of a single plant over time [152]. Another important advantage of digital imaging is that the raw data (digital images) can be easily stored and accessed later on for reanalysis, providing there is a good database in place. Obviously, adequate image analysis software is an essential feature in the concept of automated phenotyping. Several (semi-automated) image analysis tools are based on the free-ware ImageJ software, but other, more sophisticated software, for 2D or 3D analysis, were designed for measurement of leaf shape (LAMINA, [153]), root growth (GROWSCREEN\_ROOT, [154]; ROOTEDGE, [155]; RHIZOSCAN, [156]; [157]; EZ-Rhizo, [158]); and hypocotyl growth and shape (HYPOTrace, [159]). The challenge is to find the most appropriate software that can extract the information relevant for the biological questions driving the experiment. Depending on the type of digital camera and the image analysis software, different types of information that describe the phenotype of the plant can be extracted: morphological or physiological and quantitative or qualitative. For example, conventional imaging can be used to measure color, shape, length, width, and size. Video imaging or consecutive imaging of the same plant or plant part over time even allows calculating plant growth. Thermal imaging (infrared thermography) is used to measure temperature and transpiration, while fluorescence and reflectance imaging allows measuring photosynthetic activity of a plant leaf or canopy. Leaf density in the canopy can be analyzed using light detection and ranging imaging that is based on laser scanning. It is now also possible to use typical clinical imaging applications, for example, magnetic resonance imaging (MRI, alternatively called nuclear magnetic resonance) for studying plant hydraulics (water or sap flow) and carbon biomass production [160]. X-ray computed tomography and a combined analysis using MRI and positron emission tomography allow the nondestructive analysis of root growth in soil environments [154]. All these types of image-derived information can be relevant to study abiotic stress tolerance mechanisms, but temperature and photosynthetic activity of leaves are among the most widely studied phenotypes in the response to abiotic stresses.

Imaging is particularly useful for the rapid, early-stage detection of abiotic stress in a plant because it allows detecting changes in the plant performance beyond the naked eye. It is therefore a unique way to diagnose plants. Understanding how plants sense stresses can be considered as a prerequisite for engineering stress tolerance. It can be assumed that plants that have reduced stress-sensing capacities will continue to grow in mild stress conditions for longer time and thereby produce higher yields. In contrast, when subjected to severe or lethal stresses, the same plants could have a disadvantage because the investment into energy-demanding processes such as growth is high and tolerance mechanisms may be induced only late. Therefore, when exposed to severe stress conditions, plants ideally should be able to increase their stress-sensing ability.

### Future Directions

Increased protection of plants against abiotic stresses involves a complex regulatory network controlling morphological, physiological, biochemical, and molecular changes. Understanding such changes has been of key importance in breeding. Breeding crop varieties with improved performance under suboptimal growing conditions is now one of the ambitious, but crucial objectives in modern plant biotechnology. Despite the great progress in the last decade, it is now apparent that abiotic stress tolerance is a complex trait involving genes with different biological functions. However, the number of genes that were actually successfully used to improve crop performance in the field is still low. Although many types of genes or processes have been the target of genetic engineering, certain areas are still unexplored. Interesting developments were recently made in the field of miRNA regulation and epigenetics [161]. Another unexplored field is that of small peptides and their role during abiotic stress signaling.

It is very unlikely that there will be one or a few solutions for current agronomic challenges caused by abiotic stresses, but rather specific solutions for specific cases. Most of the current findings originate from work in model plant species, but much more work is needed to translate such basic research findings to crops. Because of the complex nature of abiotic stress tolerance in plants, it appears that modifying one gene to

induce stress tolerance will be in many cases not sufficient. Hence, the approach of using regulon biotechnology for crop improvement, modifying a functional process or even multiple interconnected functional processes instead of one single function, appears to be the way forward. However, this approach is often limited by the lack of knowledge on the network of molecular mechanisms underpinning stress tolerance. Conventional approaches were also hampered by experimental limitations, which can now (partially) be overcome by the advent of automated phenotyping platforms for both model and crop species. Based on the efforts made for phenotyping of *Arabidopsis* plants, it can be anticipated that this small weed will remain the primary model system of choice for academic/basic research in near future. However, “omics” platforms for crop species are becoming available and it will not take too long before research in model crop species will become intensified. Validation of results obtained in *Arabidopsis* in (model) crops will increase the value of the lead technologies and justify the large investments for field trials.

After the development of high-throughput, automated platforms to study plant phenotypes in various environments, it will be essential to combine the data obtained from the various “omics” platforms to intensify biological modeling. Modeling has proven to be useful to study specific processes, such as root growth and flowering. Efforts are now being made to model abiotic stress tolerance in plants [162]. Improved phenotyping technologies will not only allow better experimental validation of predictions drawn from such biological models, but also to improve existing models by including knowledge on the impact of external factors such as abiotic stresses. By including modeling in biological studies, our understanding of the plants’ response to abiotic stress will reach a systems biology level, a new quality in plant sciences, which will accelerate the breeding process toward abiotic stress tolerance in all our major crops.

### Bibliography

1. Fedoroff NV, Battisti DS, Beachy RN, Cooper PJ, Fischhoff DA, Hodges CN, Knauf VC, Lobell D, Mazur BJ, Molden D, Reynolds MP, Ronald PC, Rosegrant MW, Sanchez PA, Vonshak A, Zhu JK (2010) Radically rethinking agriculture for the 21st century. *Science* 327:833–834



2. Buchanan BB, Gruissem W, Jones RL (eds) (2000) Biochemistry and molecular biology of plants. American Society of Plant Physiologists, Rockville
3. Ribaut JM, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. *J Exp Bot* 58:351–360
4. Sreenivasulu N, Sopory SK, Kavi Kishor PB (2007) Deciphering the regulatory mechanisms of abiotic stress tolerance in plants by genomic approaches. *Gene* 388:1–13
5. Witcombe JR, Hollington PA, Howarth CJ, Reader S, Steele KA (2008) Breeding for abiotic stresses for sustainable agriculture. *Philos Trans R Soc Lond B Biol Sci* 363:703–716
6. Vij S, Tyagi AK (2007) Emerging trends in the functional genomics of the abiotic stress response in crop plants. *Plant Biotechnol J* 5:361–380
7. Bhatnagar-Mathur P, Vadez V, Sharma KK (2008) Transgenic approaches for abiotic stress tolerance in plants: retrospect and prospects. *Plant Cell Rep* 27:411–424
8. Shinozaki K, Yamaguchi-Shinozaki K (2007) Gene networks involved in drought stress response and tolerance. *J Exp Bot* 58:221–227
9. Divya K, Jami SK, Kirti PB (2010) Constitutive expression of mustard annexin, AnnBj1 enhances abiotic stress tolerance and fiber quality in cotton under stress. *Plant Mol Biol* 73:293–308
10. Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu Rev Plant Biol* 55:373–399
11. Laloi C, Apel K, Danon A (2004) Reactive oxygen signalling: the latest news. *Curr Opin Plant Biol* 7:323–328
12. Halliwell B (2006) Reactive species and antioxidants Redox biology is a fundamental theme of aerobic life. *Plant Physiol* 141:312–322
13. Van Breusegem F, Dat JF (2006) Reactive oxygen species in plant cell death. *Plant Physiol* 141:384–390
14. Badawi GH, Kawano N, Yamauchi Y, Shimada E, Sasaki R, Kubo A, Tanaka K (2004) Over-expression of ascorbate peroxidase in tobacco chloroplasts enhances the tolerance to salt stress and water deficit. *Physiol Plant* 121:231–238
15. Basu U, Good AG, Taylor GJ (2001) Transgenic Brassica napus plants overexpressing aluminium- induced mitochondrial manganese superoxide dismutase cDNA are resistant to aluminium. *Plant Cell Environ* 24:1269–1278
16. Gaber A, Yoshimura K, Yamamoto T, Yabuta Y, Takeda T, Miyasaka H, Nakano Y, Shigeoka S (2006) Glutathione peroxidase-like protein of Synechocystis PCC 6803 confers tolerance to oxidative and environmental stresses in transgenic Arabidopsis. *Physiol Plant* 128:251–262
17. Gong H, Jiao Y, Hu WW, Pua EC (2005) Expression of glutathione-S-transferase and its role in plant growth and development in vivo and shoot morphogenesis in vitro. *Plant Mol Biol* 57:53–66
18. Yoshimura K, Miyao K, Gaber A, Takeda T, Kanaboshi H, Miyasaka H, Shigeoka S (2004) Enhancement of stress tolerance in transgenic tobacco plants overexpressing Chlamydomonas glutathione peroxidase in chloroplasts or cytosol. *Plant J* 37:21–33
19. Murgia I, Tarantino D, Vannini C, Bracale M, Carravieri S, Soave C (2004) Arabidopsis thaliana plants overexpressing thylakoidal ascorbate peroxidase show increased resistance to Paraquat-induced photooxidative stress and to nitric oxide-induced cell death. *Plant J* 38:940–953
20. Oberschall A, Deák M, Török K, Sass L, Vass I, Kovács I, Fehér A, Dudits D, Horváth GV (2000) A novel aldose/aldehyde reductase protects transgenic plants against lipid peroxidation under chemical and drought stresses. *Plant J* 24:437–446
21. Sunkar R, Kapoor A, Zhu JK (2006) Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell* 18:2051–2065
22. Yamamoto A, Bhuiyan MN, Waditee R, Tanaka Y, Esaka M, Oba K, Jagendorf AT, Takabe T (2005) Suppressed expression of the apoplastic ascorbate oxidase gene increases salt tolerance in tobacco and Arabidopsis plants. *J Exp Bot* 56:1785–1796
23. Zhao F, Zhang H (2006) Expression of Suaeda salsa glutathione S-transferase in transgenic rice resulted in a different level of abiotic stress resistance. *J Agric Sci* 144:547–554
24. Korniyev D, Logan BA, Payton P, Allen RD, Holaday AS (2001) Enhanced photochemical light utilization and decreased chilling-induced photoinhibition of photosystem II in cotton overexpressing genes encoding chloroplast-targeted antioxidant enzymes. *Physiol Plant* 113:323–331
25. Matsumura T, Tabayashi N, Kamagata Y, Souma C, Saruyama H (2002) Wheat catalase expressed in transgenic rice can improve tolerance against low temperature stress. *Physiol Plant* 116:317–327
26. Wang FZ, Wang QB, Kwon SY, Kwak SS, Su WA (2005) Enhanced drought tolerance of transgenic rice plants expressing a pea manganese superoxide dismutase. *J Plant Physiol* 162:465–472
27. Lee SH, Ahsan N, Lee KW, Kim DH, Lee DG, Kwak SS, Kwon SY, Kim TH, Lee BH (2007) Simultaneous overexpression of both CuZn superoxide dismutase and ascorbate peroxidase in transgenic tall fescue plants confers increased tolerance to a wide range of abiotic stresses. *J Plant Physiol* 164:1626–1638
28. Prashanth SR, Sadhasivam V, Parida A (2008) Over expression of cytosolic copper/zinc superoxide dismutase from a mangrove plant Avicennia marina in indica rice var Pusa Basmati-1 confers abiotic stress tolerance. *Transgenic Res* 17:281–291
29. Huang W, Ma X, Wang Q, Gao Y, Xue Y, Niu X, Yu G, Liu Y (2008) Significant improvement of stress tolerance in tobacco plants by overexpressing a stress-responsive aldehyde dehydrogenase gene from maize (Zea mays). *Plant Mol Biol* 68:451–463
30. Xue T, Li X, Zhu W, Wu C, Yang G, Zheng C (2009) Cotton metallothionein GhMT3a, a reactive oxygen species scavenger, increased tolerance against abiotic stress in transgenic tobacco and yeast. *J Exp Bot* 60:339–349

31. Van Aken O, Giraud E, Clifton R, Whelan J (2009) Alternative oxidase: a target and regulator of stress responses. *Physiol Plant* 137:354–361
32. Zhang Y, Xi D, Wang J, Zhu D, Guo X (2009) Functional analysis reveals effects of tobacco alternative oxidase gene (NtAOX1a) on regulation of defence responses against abiotic and biotic stresses. *Biosci Rep* 29:375–383
33. Wang W, Vinocur B, Shoseyov O, Altman A (2004) Role of plant heat shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci* 9:244–252
34. Katiyar-Agarwal S, Agarwal M, Grover A (2003) Heat-tolerant basmati rice engineered by over-expression of hsp101. *Plant Mol Biol* 51:677–686
35. Guo S-J, Zhou H-Y, Zhang XS, Li X-G, Meng Q-W (2007) Overexpression of CaHSP26 in transgenic tobacco alleviates photoinhibition of PSII and PSI during chilling stress under low irradiance. *J Plant Physiol* 164:126–136
36. Hong S-W, Vierling E (2000) Mutants of *Arabidopsis thaliana* defective in the acquisition of tolerance to high temperature stress. *Proc Natl Acad Sci USA* 97:4392–4397
37. Lee JH, Schöffl F (1996) An Hsp70 antisense gene affects the expression of HSP70/HSC70, the regulation of HSF, and the acquisition of thermotolerance in transgenic *Arabidopsis thaliana*. *Mol Gen Genet* 252:11–19
38. Malik MK, Slovin JP, Hwang CH, Zimmerman JL (1999) Modified expression of a carrot small heat shock protein gene, hsp17.7, results in increased or decreased thermotolerance. *Plant J* 20:89–99
39. Park SM, Hong CB (2002) Class I small heat-shock protein gives thermotolerance in tobacco. *J Plant Physiol* 159:25–30
40. Queitsch C, Hong SW, Vierling E, Lindquist S (2000) Heat shock protein 101 plays a crucial role in thermotolerance in *Arabidopsis*. *Plant Cell* 12:479–492
41. Rhoads DM, White SJ, Zhou Y, Muralidharan M, Elthon TE (2005) Altered gene expression in plants with constitutive expression of a mitochondrial small heat shock protein suggests the involvement of retrograde regulation in the heat stress response. *Physiol Plant* 123:435–444
42. Yang JY, Sun Y, Sun AQ, Yi SY, Qin J, Li MH, Liu J (2006) The involvement of chloroplast HSP100/ClpB in the acquired thermotolerance in tomato. *Plant Mol Biol* 62:385–395
43. Maqbool A, Abbas W, Rao AQ, Irfan M, Zahur M, Bakhsh A, Riazuddin S, Husnain T (2010) *Gossypium arboreum* GHSP26 enhances drought tolerance in *Gossypium hirsutum*. *Biotechnol Prog* 26:21–25
44. Wang W, Vinocur B, Altman A (2003) Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. *Planta* 218:1–14
45. Shinozaki K, Yamaguchi-Shinozaki K (2000) Molecular response to dehydration and low temperature: Differences and cross-talk between two stress signaling pathways. *Curr Opin Plant Biol* 3:217–223
46. Zhu JK (2002) Salt and drought stress signal transduction in plants. *Annu Rev Plant Biol* 53:247–273
47. Thomashow MF (1998) Role of cold-responsive genes in plant freezing tolerance. *Plant Physiol* 118:1–8
48. Thomashow MF (1999) Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. *Annu Rev Plant Physiol Plant Mol Biol* 50:571–599
49. Xiao B, Huang Y, Tang N, Xiong L (2007) Over-expression of a LEA gene in rice improves drought resistance under the field conditions. *Theor Appl Genet* 115:35–46
50. Delauney AJ, Verma DPS (1993) Proline biosynthesis and osmoregulation in plants. *Plant J* 4:215–223
51. Tarczynski MC, Jensen RG, Bohnert HJ (1993) Stress protection of transgenic tobacco by production of the osmolyte mannitol. *Science* 259:508–510
52. Hasegawa PM, Bressan RA, Zhu JK, Bohnert HJ (2000) Plant cellular and molecular responses to high salinity. *Annu Rev Plant Physiol Plant Mol Biol* 51:463–499
53. Vinocur B, Altman A (2005) Recent advances in engineering plant tolerance to abiotic stress: achievements and limitation. *Curr Opin Biotechnol* 16:123–132
54. Kathuria H, Giri J, Nataraja KN, Murata N, Udayakumar M, Tyagi AK (2009) Glycinebetaine-induced water-stress tolerance in codA-expressing transgenic indica rice is associated with up-regulation of several stress responsive genes. *Plant Biotechnol J* 7:512–526
55. Jang IC, Oh SJ, Seo JS, Choi WB, Song SI, Kim CH, Kim YS, Seo HS, Choi YD, Nahm BH, Kim JK (2003) Expression of a bifunctional fusion of the *Escherichia coli* genes for trehalose-6-phosphate synthase and trehalose-6-phosphate phosphatase in transgenic rice plants increases trehalose accumulation and abiotic stress tolerance without stunting growth. *Plant Physiol* 131:516–524
56. Garg AK, Kim J-K, Owens TG, Ranwala AP, Choi YD, Kochian LV, Wu RJ (2002) Trehalose accumulation in rice plants confers high tolerance levels to different abiotic stresses. *Proc Natl Acad Sci USA* 99:15898–15903
57. Quan R, Shang M, Zhang H, Zhao Y, Zhang J (2004) Engineering of enhanced glycine betaine synthesis improves drought tolerance in maize. *Plant Biotechnol J* 2:477–486
58. Park EJ, Jeknić Z, Pino MT, Murata N, Chen TH (2007) Glycinebetaine accumulation is more effective in chloroplasts than in the cytosol for protecting transgenic tomato plants against abiotic stress. *Plant Cell Environ* 30:994–1005
59. Abebe T, Guenzi AC, Martin B, Cushman JC (2003) Tolerance of mannitol-accumulating transgenic wheat to water stress and salinity. *Plant Physiol* 131:1748–1755
60. Vendruscolo EC, Schuster I, Pileggi M, Scapim CA, Molinari HB, Marur CJ, Vieira LG (2007) Stress-induced synthesis of proline confers tolerance to water deficit in transgenic wheat. *J Plant Physiol* 164:1367–1376
61. Shirasawa K, Takabe T, Takabe T, Kishitani S (2006) Accumulation of glycinebetaine in rice plants that overexpress choline monooxygenase from spinach and evaluation of their tolerance to abiotic stress. *Ann Bot* 98:565–571
62. Diamant S, Eliahu N, Rosenthal D, Goloubinoff P (2001) Chemical chaperones regulate molecular chaperones in vitro and in

- cells under combined salt and heat stresses. *J Biol Chem* 276:39586–39591
63. Hare PD, Cress WA, Van Staden J (1998) Dissecting the roles of osmolyte accumulation during stress. *Plant Cell Environ* 21:535–553
  64. Tuteja N (2007) Mechanisms of high salinity tolerance in plants. *Methods Enzymol* 428:419–438
  65. Obata T, Kitamoto HK, Nakamura A, Fukuda A, Tanaka Y (2007) Rice shaker potassium channel OsKAT1 confers tolerance to salinity stress on yeast and rice cells. *Plant Physiol* 144:1978–1985
  66. Fukuda A, Chiba K, Maeda M, Nakamura A, Maeshima M, Tanaka Y (2004) Effect of salt and osmotic stresses on the expression of genes for the vacuolar H<sup>+</sup>-pyrophosphatase, H<sup>+</sup>-ATPase subunit A, and Na<sup>+</sup>/H<sup>+</sup> antiporter from barley. *J Exp Bot* 55:585–594
  67. Laurie S, Feeney KA, Maathuis FJ, Heard PJ, Brown SJ, Leigh RA (2002) A role for HKT1 in sodium uptake by wheat roots. *Plant J* 32:139–149
  68. Chung JS, Zhu JK, Bressan RA, Hasegawa PM, Shi H (2008) Reactive oxygen species mediate Na<sup>+</sup>-induced SOS1 mRNA stability in Arabidopsis. *Plant J* 53:554–565
  69. Lian HL, Yu X, Ye Q, Ding X, Kitagawa Y, Kwak SS, Su WA, Tang ZC (2004) The role of aquaporin RWC3 in drought avoidance in rice. *Plant Cell Physiol* 45:481–489
  70. Pereira JF, Zhou G, Delhaize E, Richardson T, Zhou M, Ryan PR (2010) Engineering greater aluminium resistance in wheat by over-expressing TaALMT1. *Ann Bot* 106:205–214
  71. Sasaki T, Yamamoto Y, Ezaki B, Katsuhara M, Ahn SJ, Ryan PR, Delhaize E, Matsumoto H (2004) A wheat gene encoding an aluminum-activated malate transporter. *Plant J* 37:645–653
  72. Delhaize E, Ryan PR, Hebb DM, Yamamoto Y, Sasaki T, Matsumoto H (2004) Engineering high-level aluminum tolerance in barley with the ALMT1 gene. *Proc Natl Acad Sci USA* 101:15249–15254
  73. Ariizumi T, Kishitani S, Inatsugi R, Nishida I, Murata N, Toriyama K (2002) An increase in unsaturation of fatty acids in phosphatidylglycerol from leaves improves the rates of photosynthesis and growth at low temperatures in transgenic rice seedlings. *Plant Cell Physiol* 43:751–758
  74. Khodakovskaya M, McAvoy R, Peters J, Wu H, Li Y (2006) Enhanced cold tolerance in transgenic tobacco expressing a chloroplast omega-3 fatty acid desaturase gene under the control of a cold-inducible promoter. *Planta* 223:1090–1100
  75. Orlova IV, Serebriiskaya TS, Popov V, Merkulova N, Nosov AM, Trunova TI, Tsydendambaev VD, Los DA (2003) Transformation of tobacco with a gene for the thermophilic acyl-lipid desaturase enhances the chilling tolerance of plants. *Plant Cell Physiol* 44:447–450
  76. Sui N, Li M, Zhao SJ, Li F, Liang H, Meng QW (2007) Overexpression of glycerol-3-phosphate acyltransferase gene improves chilling tolerance in tomato. *Planta* 226:1097–1108
  77. Zhang M, Barg R, Yin M, Gueta-Dahan Y, Leikin-Frenkel A, Salts Y, Shabtai S, Ben-Hayyim G (2005) Modulated fatty acid desaturation via overexpression of two distinct  $\omega$ -3 desaturases differentially alters tolerance to various abiotic stresses in transgenic tobacco cells and plants. *Plant J* 44:361–371
  78. Sohn SO, Back K (2007) Transgenic rice tolerant to high temperature with elevated contents of dienoic fatty acids. *Biol Plant* 51:340–342
  79. Osakabe Y, Mizuno S, Tanaka H, Maruyama K, Osakabe K, Todaka D, Fujita Y, Kobayashi M, Shinozaki K, Yamaguchi-Shinozaki K (2010) Overproduction of the membrane-bound receptor-like protein kinase 1, RPK1, enhances abiotic stress tolerance in Arabidopsis. *J Biol Chem* 285:9190–9201
  80. Yang L, Ji W, Zhu Y, Gao P, Li Y, Cai H, Bai X, Guo D (2010). GsCBRLK, a calcium/calmodulin-binding receptor-like kinase, is a positive regulator of plant tolerance to salt and ABA stress. *J Exp Bot* 61:2519–2533
  81. Ouyang SQ, Liu YF, Liu P, Lei G, He SJ, Ma B, Zhang WK, Zhang JS, Chen SY (2010) Receptor-like kinase OsSIK1 improves drought and salt stress tolerance in rice (*Oryza sativa*) plants. *Plant J* 62:316–329
  82. Raghavendra AS, Gonugunta VK, Christmann A, Grill E (2010) ABA perception and signaling. *Trends Plant Sci* 15:395–401
  83. Melcher K, Ng LM, Zhou XE, Soon FF, Xu Y, Suino-Powell KM, Park SY, Weiner JJ, Fujii H, Chinnusamy V, Kovach A, Li J, Wang Y, Li J, Peterson FC, Jensen DR, Yong EL, Volkman BF, Cutler SR, Zhu JK, Xu HE (2009) A gate-latch-lock mechanism for hormone signalling by abscisic acid receptors. *Nature* 462:602–608
  84. Santiago J, Dupeux F, Round A, Antoni R, Park SY, Jamin M, Cutler SR, Rodriguez PL, Márquez JA (2009) The abscisic acid receptor PYR1 in complex with abscisic acid. *Nature* 462:665–668
  85. Koh S, Lee SC, Kim MK, Koh JH, Lee S, An G, Choe S, Kim SR (2007) T-DNA tagged knockout mutation of rice OsGSK1, an orthologue of Arabidopsis BIN2, with enhanced tolerance to various abiotic stresses. *Plant Mol Biol* 65:453–466
  86. Diédhiou CJ, Popova OV, Dietz KJ, Gollack D (2008) The SNF1-type serine-threonine protein kinase SAPK4 regulates stress-responsive gene expression in rice. *BMC Plant Biol* 28(8):49
  87. Saijo Y, Hata S, Kyojuka J, Shimamoto K, Izui K (2000) Overexpression of a single Ca<sup>2+</sup>-dependent protein kinase confers both cold and salt/drought tolerance on rice plants. *Plant J* 23:319–327
  88. Xiang Y, Huang Y, Xiong L (2007) Characterization of stress-responsive CIPK genes in rice for stress tolerance improvement. *Plant Physiol* 144:1416–1428
  89. Shou H, Bordallo P, Wang K (2004) Expression of the Nicotiana protein kinase (NPK1) enhanced drought tolerance in transgenic maize. *J Exp Bot* 55:1013–1019
  90. Agarwal PK, Agarwal P, Jain P, Jha B, Reddy MK, Sopory SK (2008) Constitutive overexpression of a stress-inducible small GTP-binding protein PgRab7 from *Pennisetum glaucum* enhances abiotic stress tolerance in transgenic tobacco. *Plant Cell Rep* 27:105–115
  91. Seki M, Kamei A, Yamaguchi-Shinozaki K, Shinozaki K (2003) Molecular responses to drought, salinity and frost: common and different paths for plant protection. *Curr Opin Biotechnol* 14:194–199

92. Shinozaki K, Yamaguchi-Shinozaki K, Seki M (2003) Regulatory network of gene expression in the drought and cold stress responses. *Curr Opin Plant Biol* 6:410–417
93. Nakashima K, Yamaguchi-Shinozaki K (2005) Molecular studies on stress-responsive gene expression in Arabidopsis and improvement of stress tolerance in crop plants by regulon biotechnology. *Jpn Agr Res Q* 39:221–229
94. Umezawa T, Fujita M, Fujita Y, Yamaguchi-Shinozaki K, Shinozaki K (2006) Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. *Curr Opin Biotechnol* 17:113–122
95. Agarwal PK, Agarwal P, Reddy MK, Sopory SK (2006) Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. *Plant Cell Rep* 25:1263–1274
96. Yamaguchi-Shinozaki K, Shinozaki K (2006) Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu Rev Plant Biol* 57:781–803
97. Ito Y, Katsura K, Maruyama K, Taji T, Kobayashi M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2006) Functional analysis of rice DREB1/CBF-type transcription factors involved in cold-responsive gene expression in transgenic rice. *Plant Cell Physiol* 47:141–153
98. Savitch LV, Allard G, Seki M, Robert LS, Tinker NA, Huner NP, Shinozaki K, Singh J (2005) The effect of overexpression of two Brassica CBF/DREB1-like transcription factors on photosynthetic capacity and freezing tolerance in Brassica napus. *Plant Cell Physiol* 46:1525–1539
99. Nelson DE, Repetti PP, Adams TR, Creelman RA, Wu J, Warner DC, Anstrom DC, Bensen RJ, Castiglioni PP, Donnarummo MG, Hinchey BS, Kumimoto RW, Maszle DR, Canales RD, Krolkowski KA, Dotson SB, Gutterson N, Ratcliffe OJ, Heard JE (2007) Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc Natl Acad Sci USA* 104:16450–16455
100. Lee HE, Shin D, Park SR, Han SE, Jeong MJ, Kwon TR, Lee SK, Park SC, Yi BY, Kwon HB, Byun MO (2007) Ethylene responsive element binding protein 1 (StEREBP1) from Solanum tuberosum increases tolerance to abiotic stress in transgenic potato plants. *Biochem Biophys Res Commun* 353:863–868
101. Zhang G, Chen M, Li L, Xu Z, Chen X, Guo J, Ma Y (2009) Overexpression of the soybean GmERF3 gene, an AP2/ERF type transcription factor for increased tolerances to salt, drought, and diseases in transgenic tobacco. *J Exp Bot* 60:3781–3796
102. Karaba A, Dixit S, Greco R, Aharoni A, Trijatmiko KR, Marsch-Martinez N, Krishnan A, Nataraja KN, Udayakumar M, Pereira A (2007) Improvement of water use efficiency in rice by expression of HARDY, an Arabidopsis drought and salt tolerance gene. *Proc Natl Acad Sci USA* 104:15270–15275
103. Gao S, Zhang H, Tian Y, Li F, Zhang Z, Lu X, Chen X, Huang R (2008) Expression of TERF1 in rice regulates expression of stress-responsive genes and enhances tolerance to drought and high-salinity. *Plant Cell Rep* 27:1787–1795
104. Xiang Y, Tang N, Du H, Ye H, Xiong L (2008) Characterization of OsbZIP23 as a key player of the basic leucine zipper transcription factor family for conferring abscisic acid sensitivity and salinity and drought tolerance in rice. *Plant Physiol* 148:1938–1952
105. Huang J, Sun SJ, Xu DQ, Yang X, Bao YM, Wang ZF, Tang HJ, Zhang H (2009) Increased tolerance of rice to cold, drought and oxidative stresses mediated by the overexpression of a gene that encodes the zinc finger protein ZFP245. *Biochem Biophys Res Commun* 389:556–561
106. Ye H, Du H, Tang N, Li X, Xiong L (2009) Identification and expression profiling analysis of TIFY family genes involved in stress and phytohormone responses in rice. *Plant Mol Biol* 71:291–305
107. Ma Q, Dai X, Xu Y, Guo J, Liu Y, Chen N, Xiao J, Zhang D, Xu Z, Zhang X, Chong K (2009) Enhanced tolerance to chilling stress in OsMYB3R-2 transgenic rice is mediated by alteration in cell cycle and ectopic expression of stress genes. *Plant Physiol* 150:244–256
108. Ogo Y, Itai RN, Nakanishi H, Kobayashi T, Takahashi M, Mori S, Nishizawa NK (2007) The rice bHLH protein OsIRO2 is an essential regulator of the genes involved in Fe uptake under Fe-deficient conditions. *Plant J* 51:366–377
109. Nakashima K, Tran LS, Van Nguyen D, Fujita M, Maruyama K, Todaka D, Ito Y, Hayashi N, Shinozaki K, Yamaguchi-Shinozaki K (2007) Functional analysis of a NAC-type transcription factor OsNAC6 involved in abiotic and biotic stress-responsive gene expression in rice. *Plant J* 51:617–630
110. Yi SY, Kim JH, Joung YH, Lee S, Kim WT, Yu SH, Choi D (2004) The pepper transcription factor CaPF1 confers pathogen and freezing tolerance in Arabidopsis. *Plant Physiol* 136:2862–2874
111. Hu H, Dai M, Yao J, Xiao B, Li X, Zhang Q, Xiong L (2006) Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. *Proc Natl Acad Sci USA* 103:12987–12992
112. Huang XY, Chao DY, Gao JP, Zhu MZ, Shi M, Lin HX (2009) A previously unknown zinc finger protein, DST, regulates drought and salt tolerance in rice via stomatal aperture control. *Genes Dev* 23:1805–1817
113. Vannini C, Campa M, Iriti M, Genga A, Faoro F, Carravieri S, Rotino GL, Rossoni M, Spinardi AM (2007) Evaluation of transgenic tomato plants ectopically expressing the rice Osmyb4 gene. *Plant Sci* 173:231–239
114. Seong ES, Baek K-H, Oh S-K, Jo SH, Yi SY, Park JM, Joung YH, Lee S, Cho HS, Choi D (2007) Induction of enhanced tolerance to cold stress and disease by overexpression of the pepper CaPIF1 gene in tomato. *Physiol Plant* 129:555–566
115. Hsieh TH, Li CW, Su RC, Cheng CP, Sanjaya TYC, Chan MT (2010) A tomato bZIP transcription factor, SIAREB, is involved in water deficit and salt stress response. *Planta* 231:1459–1473
116. Guo YH, Yu YP, Wang D, Wu CA, Yang GD, Huang JG, Zheng CC (2009) GhZFP1, a novel CCCH-type zinc finger protein from cotton, enhances salt stress tolerance and fungal disease resistance in transgenic tobacco by interacting with GZIRD21A and GZIPR5. *New Phytol* 183:62–75

117. Huang J, Wang MM, Jiang Y, Bao YM, Huang X, Sun H, Xu DQ, Lan HX, Zhang HS (2008) Expression analysis of rice A20/AN1-type zinc finger genes and characterization of ZFP177 that contributes to temperature stress tolerance. *Gene* 420:135–144
118. Shukla LI, Chinnusamy V, Sunkar R (2008) The role of microRNAs and other endogenous small RNAs in plant stress responses. *Biochim Biophys Acta* 1779:743–748
119. Mittler R, Feng X, Cohen M (1998) Post-transcriptional suppression of cytosolic ascorbate peroxidase expression during pathogen-induced programmed cell death in tobacco. *Plant Cell* 10:461–473
120. Mittler R, Zilinskas BA (1994) Regulation of pea cytosolic ascorbate peroxidase and other antioxidant enzymes during the progression of drought stress and following recovery from drought. *Plant J* 5:397–405
121. Forment J, Naranjo MA, Roldán M, Serrano R, Vicente O (2002) Expression of Arabidopsis SR-like splicing proteins confers salt tolerance to yeast and transgenic plants. *Plant J* 30:511–519
122. Kant P, Kant S, Gordon M, Shaked R, Barak S (2007) STRESS RESPONSE SUPPRESSOR1 and STRESS RESPONSE SUPPRESSOR2, two DEAD-box RNA helicases that attenuate Arabidopsis responses to multiple abiotic stresses. *Plant Physiol* 145:814–830
123. Gong Z, Dong CH, Lee H, Zhu J, Xiong L, Gong D, Stevenson B, Zhu JK (2005) A DEAD box RNA helicase is essential for mRNA export and important for development and stress responses in Arabidopsis. *Plant Cell* 17:256–267
124. Kim YO, Kim JS, Kang H (2005) Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in Arabidopsis thaliana. *Plant J* 42:890–900
125. Kim YO, Pan S, Jung C-H, Kang H (2007) A zinc finger-containing glycine-rich RNA-binding protein, at RZ-1a, has a negative impact on seed germination and seedling growth of Arabidopsis thaliana under salt or drought stress conditions. *Plant Cell Physiol* 48:1170–1181
126. Romero C, Belles JM, Vaya JL, Serrano R, Culianez-Macia FA (1997) Expression of the yeast trehalose-6-phosphate synthase gene in transgenic tobacco plants: pleiotropic phenotypes include drought tolerance. *Planta* 201:293–297
127. Zhang Y, Yang C, Li Y, Zheng N, Chen H, Zhao Q, Gao T, Guo H, Xie Q (2007) SDIR1 is a RING finger E3 ligase that positively regulates stress-responsive abscisic acid signaling in Arabidopsis. *Plant Cell* 19:1912–1929
128. Zhang YY, Li Y, Gao T, Zhu H, Wang DJ, Zhang HW, Ning YS, Liu LJ, Wu YR, Chu CC, Guo HS, Xie Q (2008) Arabidopsis SDIR1 enhances drought tolerance in crop plants. *Biosci Biotechnol Biochem* 72:2251–2254
129. Guo X, Wu Y, Wang Y, Chen Y, Chu C (2009) OsMSRA4.1 and OsMSRB1.1, two rice plastidial methionine sulfoxide reductases, are involved in abiotic stress responses. *Planta* 230:227–238
130. Wang Y, Ying J, Kuzma M, Chalifoux M, Sample A, McArthur C, Uchacz T, Sarvas C, Wan J, Dennis DT, McCourt P, Huang Y (2005) Molecular tailoring of farnesylation for plant drought tolerance and yield protection. *Plant J* 43:413–424
131. De Block M, Verduyn C, De Brouwer D, Cornelissen M (2005) Poly(ADP-ribose) polymerase in plants affects energy homeostasis, cell death and stress tolerance. *Plant J* 41:95–106
132. Vanderauwera S, De Block M, Van de Steene N, van de Cotte B, Metzlauff M, Van Breusegem F (2007) Silencing of poly(ADP-ribose) polymerase in plants alters abiotic stress signal transduction. *Proc Natl Acad Sci USA* 104:15150–15155
133. Hauben M, Haesendonckx B, Standaert E, Van Der Kelen K, Azmi A, Akpo H, Van Breusegem F, Guisez Y, Bots M, Lambert B, Laga B, De Block M (2009) Energy use efficiency is characterized by an epigenetic component that can be directed through artificial selection to increase yield. *Proc Natl Acad Sci USA* 106:20109–20114
134. Mittler R, Blumwald E (2010) Genetic engineering for modern agriculture: challenges and perspectives. *Annu Rev Plant Biol* 61:443–462
135. Yamaguchi T, Blumwald E (2005) Developing salt-tolerant crop plants: challenges and opportunities. *Trends Plant Sci* 10:615–620
136. Mittler R (2006) Abiotic stress, the field environment and stress combination. *Trends Plant Sci* 11:15–19
137. Pino MT, Skinner JS, Park EJ, Jeknić Z, Hayes PM, Thomashow MF, Chen TH (2007) Use of a stress inducible promoter to drive ectopic AtCBF expression improves potato freezing tolerance while minimizing negative effects on tuber yield. *Plant Biotechnol J* 5:591–604
138. Massonet C, Vile D, Fabre J, Hannah MA, Caldana C, Lisec J, Beemster GT, Meyer RC, Messerli G, Gronlund JT, Perkovic J, Wigmore E, May S, Bevan MW, Meyer C, Rubio-Diaz S, Weigel D, Micol JL, Buchanan-Wollaston V, Fiorani F, Walsh S, Rinn B, Grissem W, Hilson P, Hennig L, Willmitzer L, Granier C (2010) *Plant Physiol* 152:2142–2157
139. Verslues PE, Agarwal M, Katiyar-Agarwal S, Zhu J, Zhu JK (2006) Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. *Plant J* 45:523–539
140. Salekdeh GH, Reynolds M, Bennett J, Boyer J (2009) Conceptual framework for drought phenotyping during molecular breeding. *Trends Plant Sci* 14:488–496
141. Kolukisaoglu Ü, Thurow K (2010) Future and frontiers of automated screening in plant sciences. *Plant Sci* 178:476–484
142. Urano K, Kurihara Y, Seki M, Shinozaki K (2010) 'Omics' analyses of regulatory networks in plant abiotic stress responses. *Curr Opin Plant Biol* 13:132–138
143. Mochida K, Shinozaki K (2010) Genomics and bioinformatics resources for crop improvement. *Plant Cell Physiol* 51:497–523
144. Finkel E (2009) Imaging. With 'phenomics', plant scientists hope to shift breeding into overdrive. *Science* 325:380–381
145. Saint Pierre C, Crossa J, Manes Y, Reynolds MP (2010) Gene action of canopy temperature in bread wheat under diverse environments. *Theor Appl Genet* 120:1107–1117
146. Granier C, Aguirrezabal L, Chenu K, Cookson SJ, Dauzat M, Hamard P, Thioux JJ, Rolland G, Bouchier-Combaud S,



- Lebaudy A, Muller B, Simonneau T, Tardieu F (2006) PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in *Arabidopsis thaliana* permitted the identification of an accession with low sensitivity to soil water deficit. *New Phytol* 169:623–635
147. Sadok W, Naudin P, Boussuge B, Muller B, Welcker C, Tardieu F (2007) Leaf growth rate per unit thermal time follows QTL-dependent daily patterns in hundreds of maize lines under naturally fluctuating conditions. *Plant Cell Environ* 30:135–146
148. Walter A, Scharf H, Gilmer F, Zierer R, Nagel KA, Ernst M, Wiese A, Virnich O, Christ MM, Uhlig B, Jünger S, Schurr U (2007) Dynamics of seedling growth acclimation towards altered light conditions can be quantified via GROWSCREEN: a setup and procedure designed for rapid optical phenotyping of different plant species. *New Phytol* 174:447–455
149. Reuzeau C, Pen J, Frankard V, de Wolf J, Peerbolte R, Broekaert W, van Camp W (2005) TraitMill: a discovery engine for identifying yield-enhancement genes in cereals. *Mol Plant Breed* 3:753–759
150. Gregory PJ, Hutchison DJ, Read DB, Jenneson PM, Gilboy WB, Morton EJ (2003) Non-invasive imaging of roots with high resolution X-ray micro-tomography. *Plant Soil* 255:351–359
151. Yazdanbakhsh N, Fisahn J (2010) Analysis of *Arabidopsis thaliana* root growth kinetics with high temporal and spatial resolution. *Ann Bot* 105:783–791
152. Chaerle L, Lenk S, Leinonen I, Jones HG, Van Der Straeten D, Buschmann C (2009) Multi-sensor plant imaging: towards the development of a stress-catalogue. *Biotechnol J* 4:1152–1167
153. Bylesjö M, Segura V, Soolanayakanahally RY, Rae AM, Trygg J, Gustafsson P, Jansson S, Street NR (2008) LAMINA: a tool for rapid quantification of leaf size and shape parameters. *BMC Plant Biol* 22(8):82
154. Nagel KA, Kastenholz B, Jahnke S, van Dusschoten D, Aach T, Mühlich M, Truhn D, Scharf H, Terjung S, Walter A, Schurr U (2009) Temperature responses of roots: impact on growth, root system architecture and implications for phenotyping. *Funct Plant Biol* 36:947–959
155. Kasper TC, Ewing RP (1997) ROOTEDGE: software for measuring root length from desktop scanner images. *Agron J* 89:932–940
156. Berzin I, Cohen B, Mills D, Dinstein I, Merchuk JC (2000) RHIZOSCAN: a semiautomatic image processing system for characterization of the morphology and secondary metabolite concentration in hairy root cultures. *Biotechnol Bioeng* 70:17–24
157. French A, Ubeda-Tomás S, Holman TJ, Bennett MJ, Pridmore T (2009) High-throughput quantification of root growth using a novel image-analysis tool. *Plant Physiol* 150:1784–1795
158. Armengaud P (2009) EZ-Rhizo software: the gateway to root architecture analysis. *Plant Signal Behav* 4:139–141
159. Wang L, Uilecan IV, Assadi AH, Kozmik CA, Spalding EP (2009) HYPOTrace: image analysis software for measuring hypocotyl growth and shape demonstrated on *Arabidopsis* seedlings undergoing photomorphogenesis. *Plant Physiol* 149:1632–1637
160. Van As H (2007) Intact plant MRI for the study of cell water relations, membrane permeability, cell-to-cell and long distance water transport. *J Exp Bot* 58:743–756
161. Jia X, Mendu V, Tang G (2010) An array platform for identification of stress-responsive MicroRNAs in plants. *Methods Mol Biol* 639:253–269
162. Tardieu F, Tuberosa R (2010) Dissection and modelling of abiotic stress tolerance in plants. *Curr Opin Plant Biol* 13:206–212

## AC Machines: Permanent Magnet Synchronous and Induction Machines

A. H. RANJBAR<sup>1</sup>, BABAK FAHIMI<sup>2</sup>

<sup>1</sup>Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran

<sup>2</sup>University of Texas, Arlington, TX, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Permanent Magnet Motor Drives

AC Induction Machines

Fundamental Equations of Induction Machine

Power Electronic Inverter

IM Drive in Synchronous Reference Frame

Fundamental Equations of the IM Drive Under

Field-Oriented Control

Field-Oriented Control Implementations

Future Directions

Bibliography

### Glossary

**Permanent magnet synchronous machine (PMSM)** PMSM is an electromechanical energy conversion device that forms as an important category of electric machines. This machine comprises an AC stator and a brushless rotor that houses permanent magnet. Since the frequency of the excitation in the stator winding is proportional to the mechanical speed of the rotor, this family of machines is called “synchronous.” PMSM is being



considered for variety of applications including electric propulsion of modern vehicles.

**Surface-mounted PMSM (SM-PMSM)** A type of PMSM where permanent magnets are placed on the surface of the rotor, making it easy to build, and skewed poles are easily magnetized on this surface-mounted type to minimize cogging torque.

**Interior permanent magnet (IPM)** A type of PMSM where, unlike SM-PMSM, permanent magnets are mounted such that a magnetic saliency is established.

**Induction machine (IM)** A type of electrical machines that operates based on the principle of induction. While stator is formed by a multiphase AC winding, the rotor houses is (a) a series of short-circuited bars in the form of a squirrel cage or (b) a multiphase AC winding. The latter rotor form will allow for external injection of the current. In this family of electric machines, speed of the rotor is not an integer multiple of the stator frequency, and as such, they are named asynchronous machines.

**Permanent magnet material** Permanent magnet used in PMSM is a piece of magnetic material which, once magnetized or “charged” by an external magnetic field, retains a usefully large magnetic moment after the magnetizing force is removed. Thus, a permanent magnet becomes a source of magnetic field in the rotor.

**Field-oriented control** Field-oriented control was introduced in the beginning of 1970s. It demonstrated that an induction motor or synchronous motor could be controlled similar to that of a separately excited DC motor by the orientation of the stator mmf or current vector in relation to the rotor flux to achieve a fast dynamic response in the electromagnetic torque.

## Definition of the Subject

Electrical machines can be viewed as electromechanical energy converters where at least two magnetic fields interact with each other to produce torque. There are different types of electrical machines including: brushed DC machines, brushless DC machines, permanent magnet synchronous machines (PMSM), AC induction machines (IM), switched reluctance machines (SRM), and etc.

In 1886, the first practical DC motor was invented by Frank Julian Sprague. It was a motor capable of constant speed under variable loads. Brushes were used inevitably for this family of electric machinery which played a dominant role for a good part of the nineteenth century and early decades of the twentieth century.

In 1882, Nikola Tesla discovered the rotating magnetic field and pioneered the use of a rotary field of force to operate machines. Later on, he suggested that the commutators could be removed from a machine, and the device could operate on a rotary field of force. In 1888, Nikola Tesla invented the first AC motor, and, consequently, the polyphase power transmission system was developed. With the advent of AC machine and the effective use of the rotating magnetic field, development of brushless electric machinery attracted significant attention.

This chapter is dedicated to analysis of AC machines including induction machines (IM) and permanent magnet synchronous machines (PMSM). Both IM and PMSM take advantage of the rotating magnetic field that is produced by their multiphase stator windings. Permanent magnet synchronous machines (PMSM) also known as brushless DC machines (BLDC) form an important group of electric machinery that are being considered for variety of applications including electric propulsion of modern vehicles. Due to their high efficiency, compact structure, high torque-to-current and torque-to-volume ratios, fast dynamic response, simple mechanical construction, absence of moving contact, and easy maintenance, PMSMs are widely used in many industrial applications. However, induction motors are now the preferred choice for industrial motors due to their rugged construction, absence of brushes compared to DC machines, absence of permanent magnets (PM) compared to PMSM, and the simplicity of speed control. Induction motor is considered most suitable for almost every application because of its robust and rugged construction and reliability. Moreover, its maintenance is easy, and costs are relatively lesser. It has widespread use in industrial, commercial, and domestic applications.

## Introduction

Electromechanical energy conversion can be viewed as an interaction between at least two magnetic fields.

These fields can be constituted by (1) electric current, (2) induced magnetization in ferromagnetic materials, and (3) permanent magnets. To sustain a productive electromechanical energy conversion, the following requirements should be met:

1. At least one of the magnetic fields must be caused by a controllable current.
2. The two interacting magnetic fields should be perpendicular to each other. They can be stationary in space (i.e., DC-brushed machines) or rotating together at the same angular speed (i.e., AC machines).

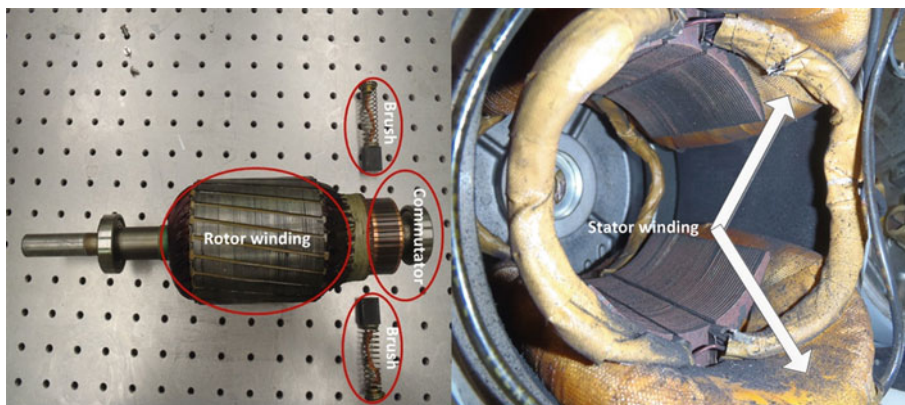
In order to understand the operation of AC machines, a quick review of DC-brushed machines is helpful. Rotor and stator of a typical two-pole DC machine are shown in Fig. 1. To compensate for the motion of the rotor, the resulting field of the rotor should be continuously kept in a stationary location. This is accomplished by the virtue of the mechanical commutator. Field of the stator can be produced by permanent magnets or fixed coils which are fed with DC current. Injection of the current to a moving member requires a continuous contact between the armature and the external supply of the direct current. This makes the use of brushes inevitable for this family of electric machinery which played a dominant role for a good part of the nineteenth century and early decades of the twentieth century.

With the advent of induction machine and the effective use of the rotating magnetic field,

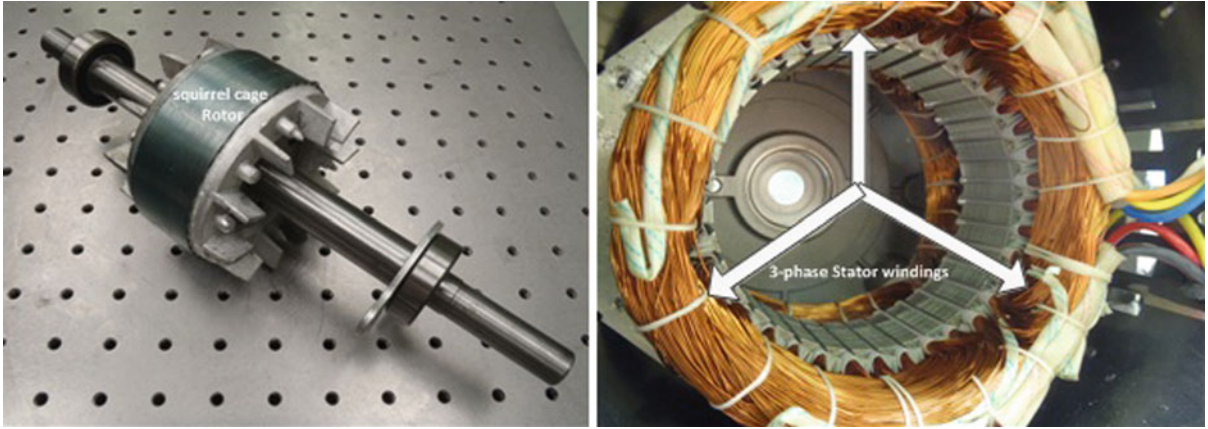
development of brushless electric machinery attracted significant attention. However, ease of control has made DC machines more popular in applications that require a wide range of motor speeds. DC-brushed machines are still among the most popular electromechanical energy converters especially for low powers where the use of permanent magnets is prominent.

This chapter is dedicated to the analysis of AC machines including induction machine (IM) and permanent magnet synchronous machine (PMSM) that are shown in Figs. 2 and 3, respectively. As shown in Figs. 2 and 3, both IM and PMSM take advantage of the rotating magnetic field that is produced by their three-phase stator windings.

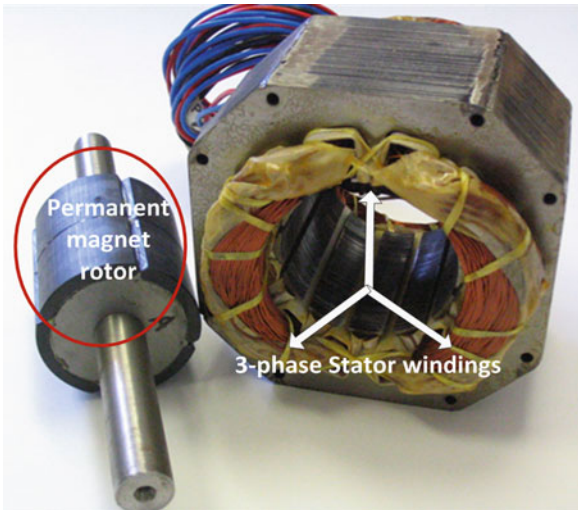
Since the magnetic field caused by the multiphase stator windings rotates with a frequency equal to supply frequency, the corresponding field of the rotor has to maintain an adequate speed and orientation to establish a relative stationary displacement with respect to the field of the stator. This can be established by rotating the rotor at a speed whose addition to the supply frequency of the rotor field (at stand still) equates that of the stator supply. While this mechanism automatically happens in a squirrel-cage induction machine, special tuning of the rotor currents in doubly fed electric machinery turns into a necessity which is properly handled by using power electronic converters. Following this design philosophy, one can intuitively expect that existence of a DC field on the rotor necessitates a mechanical speed which is proportional to that of the stator frequency. In a two-pole machine, if two



AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 1  
Rotor and stator of a two-pole DC machine



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 2**  
Stator and rotor of a squirrel-cage induction machine



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 3**  
Stator and rotor of a three-phase four-pole surface-mounted permanent magnet synchronous machine

sources of magnetic field (i.e., electric currents) with frequencies of  $f_1$  and  $f_2$  rotate at angular speeds of  $\omega_{m1}$  and  $\omega_{m2}$  respectively, then the following relationship must be held at all time to maintain synchronism between the two fields (i.e.,  $f_1 + \frac{\omega_{m1}}{2\pi} = f_2 + \frac{\omega_{m2}}{2\pi}$ ).

Employment of permanent magnets as the source of DC magnetic field on the rotor elevates the use of slip rings and brushes which are otherwise used for injection of current in a rotating circuit, hence, resulting in

a brushless configuration. Low conductivity of the permanent magnets will prohibit the appearance of appreciable induced currents at low speeds and effectively eliminates the need for providing reactive power for magnetization of the magnetic circuit. These attributes have turned permanent magnet synchronous machines into attractive alternatives for a wide range of applications with power ranging from a few watts into several megawatts.

Yet cooling, mechanical integrity, and operation under harsh ambient conditions have been reportedly impeding a universal takeover of the PMSM drives. Besides, the issue of supply chain and lower efficiency at higher speeds when field weakening has to be done are among other reasons which have raised concern over the use of PMSM in even larger scales. Yet, PMSM drives have been considered as main contenders for many applications among which electric propulsion of modern electrified automobiles, advanced renewable energy harvesting apparatus, and high-performance precision equipments stand out. The compact, relatively low-cost, and optimally sized power electronic converters pave the way for the massive integration of PMSM drives for many applications. The invention of affordable temperature-resilient high-energy permanent magnet materials (rare-earth permanent magnet) has undoubtedly accelerated the integration of PMSM into many industries.

The continued growth of automotive market for PMSM, in large part, depends on efficient operation

in the field-weakening region. The existence of permanent magnets as a steady source of magnetic field can potentially lead to supply of fault currents even when the source is disconnected leading to unsafe operating conditions. For the foreseeable future, PMSM drives continue to enjoy the affordable cost of high-energy permanent magnets and research and development will continue on development of novel magnetic configurations which can increase the efficiency of the electromechanical energy conversion and improve the reliability of this family of actuators.

### Rotating Magnetic Field Versus Stationary Magnetic Field

As mentioned earlier, both IM and PMSM contain multiphase stator windings. These multiphase stator windings produce a rotating magnetic field. In this section, a three-phase arrangement for the stator winding is considered although the same conclusion can be derived for higher number of stator phases. The rotating magnetic field produced by the three-phase AC windings in IM and PMSM can also be produced by a set of two-phase AC windings that is stationary and  $90^\circ$  apart with respect to each other. This rotating field can be symbolically represented as:

$$MMF_s = k \cos(\omega_s t - \phi_s) \quad (1)$$

In which  $\omega_s$  and  $\phi_s$  denote the frequency of current and displacement of an arbitrary point on the stator. If the frame of reference is rotating at a speed of  $\omega_{m1}$  then for an observer located at another frame of reference which has a mechanical speed of  $\omega_{m2}$  (in the same direction), the rotating field will be seen as:

$$MMF_s = k \cos((\omega_s + \omega_{m1} - \omega_{m2})t - \phi_s) \quad (2)$$

This suggests that if the stator is excited with a frequency of  $\omega_s$ , an observer in the air gap will sense the rotating field at a speed of  $\omega_s$ ; while for an observer on the rotor (rotor rotates with a speed of  $\omega_m$  in counterclockwise direction), the sensed frequency of the rotating field due to stator excitation will be  $\omega_s - \omega_m$ . Notably, the two-phase AC windings mentioned above can be replaced by two-phase DC windings that are rotating (at an angular speed equal to the AC supply) and are  $90^\circ$  apart with respect to each other. Therefore, instead of analyzing the magnetic field

produced by the stationary three-phase AC windings, one can imagine the rotating two-phase DC windings that are  $90^\circ$  apart with respect to each other. This can be a physical interpretation of Park's transformation that is a mathematical transformation used to simplify the analysis of three-phase systems.

In the case of balanced three-phase systems, the application of Park's transformation reduces the three AC quantities to two DC quantities. Simplified calculations can then be carried out on these transformed quantities before performing the inverse transform to recover the actual three-phase AC quantities. This is often used in order to simplify the analysis of three-phase synchronous machines and to simplify calculations for the control of three-phase inverters. The unique characteristic of Park's transformation is that it eliminates all time-varying inductances from the voltage equations of the machine. These time-varying inductances can originate from motion or varying magnetic reluctance. Park's transformation is expressed by Eq. 3 where  $f$  can be voltage, current, or flux linkage of a synchronous machine. Park's transformation is a special case of the reference-frame theory (also referred to as direct-quadrature ( $dq0$ ) transformation). In the case of balanced three-phase circuits, application of the  $dq0$  transform reduces the three AC quantities in the  $abc$  frame of references to two DC quantities in  $dq$  frame of references.

$$f_{dq0s} = K_S(\theta) \cdot f_{abcs} \quad (3)$$

where

$$(f_{dq0s})^T = [f_{qs} \quad f_{ds} \quad f_{os}] \quad (4)$$

$$(f_{abcs})^T = [f_{as} \quad f_{bs} \quad f_{cs}] \quad (5)$$

$$K_S(\theta) = \frac{2}{3} \begin{bmatrix} \cos \theta & \cos(\theta - \frac{2\pi}{3}) & \cos(\theta + \frac{2\pi}{3}) \\ \sin \theta & \sin(\theta - \frac{2\pi}{3}) & \sin(\theta + \frac{2\pi}{3}) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad (6)$$

$$\omega = \frac{d\theta}{dt} \quad (7)$$

In Eq. 7,  $\omega$  and  $\theta$  are supply frequency and rotor position, respectively. Thus, the definite integral form Eq. 7 can be written as Eq. 8, where  $\theta(0)$  is the initial rotor position.



$$\theta = \int_0^t \omega(\xi) d\xi + \theta(0) \quad (8)$$

It can be shown that for the inverse transformation, the following form can be used:

$$(K_S)^{-1} = \frac{2}{3} \begin{bmatrix} \cos \theta & \sin \theta & 1 \\ \cos(\theta - \frac{2\pi}{3}) & \sin(\theta - \frac{2\pi}{3}) & 1 \\ \cos(\theta + \frac{2\pi}{3}) & \sin(\theta + \frac{2\pi}{3}) & 1 \end{bmatrix} \quad (9)$$

### Torque Production in Electrical Machines

The magnetic field of stator in both IM and PMSM is produced by the virtue of three-phase stator windings which are distributed  $120^\circ$  apart from each other [1]. However, the source of magnetic field for the rotor in PMSM is different from that of IM. Assuming the squirrel-cage IM, as shown in Fig. 2, the magnetic field of the rotor originates from the voltage that is induced on the rotor bars. Since these bars are short circuited through the end ring, the voltage induced on their terminals translates into significant currents that originate the magnetic field of the rotor. Consequently, the magnetic field of the stator interacts with the magnetic field produced by rotor currents to produce torque in IM. It must be kept in mind that the resulting field of the rotor rotates at the synchronous speed in the air gap.

Unlike magnetic field of the rotor in IM that is dependent on the voltage induced by stator windings on the rotor, magnetic field of PMSM is generated by permanent magnets that are placed on the rotor. As a result, the electromagnetic torque in the PMSM results from interaction between the magnetic field of the stator (i.e., an electromagnet) and the magnetic field of the permanent magnets on the rotor. The electromagnetic torque that is the cause of rotation in electrical machines can also be defined in terms of electromagnetic co-energy. The electromagnetic co-energy that is stored in the coupling field of unsaturated electrical machines can be expressed as Eq. 10.

$$W^c = \Lambda^t(\theta_r).I \quad (10)$$

where

$$I = \begin{bmatrix} I_S \\ I_R \end{bmatrix} \quad (11)$$

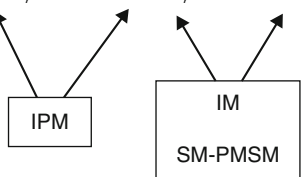
$$\Lambda = \begin{bmatrix} \Lambda_S(\theta_r) \\ \Lambda_R(\theta_r) \end{bmatrix} = \begin{bmatrix} \Lambda_{SS}(\theta_r) + \Lambda_{SR}(\theta_r) \\ \Lambda_{RR}(\theta_r) + \Lambda_{RS}(\theta_r) \end{bmatrix} \quad (12)$$

The electromagnetic torque is defined as the derivative of electromagnetic energy with respect to rotor position and is expressed by Eq. 13.

$$T = \frac{\partial W^c}{\partial \theta_r} = \frac{\partial \Lambda_{SS}^t}{\partial \theta_r} \cdot I_S + \frac{\partial \Lambda_{SR}^t}{\partial \theta_r} \cdot I_S + \frac{\partial \Lambda_{RR}^t}{\partial \theta_r} \cdot I_R + \frac{\partial \Lambda_{RS}^t}{\partial \theta_r} \cdot I_R \quad (13)$$

Based on Eq. 13, in order to have nonzero torque, the electromagnetic energy in the coupling field of the electrical machine should vary with respect to rotor position. Considering Eq. 10, since rotor and stator currents are not dependent on the rotor position, the electromagnetic torque can only be produced whenever the self-/mutual inductance of the windings changes with respect to rotor position. Therefore, the electromagnetic torque can be expressed as Eq. 14.

$$T = \frac{\partial W_f}{\partial \theta_r} = \frac{1}{2} I^T \cdot \frac{\partial}{\partial \theta_r} \begin{bmatrix} L_{SS} & L_{SR} \\ L_{RS} & L_{RR} \end{bmatrix} \cdot I \quad (14)$$

$$T = \frac{1}{2} \left\{ I_s^t \frac{\partial L_{ss}(\theta_r)}{\partial \theta_r} I_s + I_r^t \frac{\partial L_{rs}(\theta_r)}{\partial \theta_r} I_s + I_s^t \frac{\partial L_{sr}(\theta_r)}{\partial \theta_r} I_r + I_r^t \frac{\partial L_{rr}(\theta_r)}{\partial \theta_r} I_r \right\} \quad (15)$$


In Eq. 15, the torque that is produced by change of self-inductance with respect to rotor position is called “reluctance torque,” and the torque that is produced by change of mutual inductance with respect to rotor position is called “reaction torque.” Therefore, depending on the specific structure of the machine, any of these two torque components or both of them may exist. Considering the rotor and stator structures in the IM and PMSM, as shown in Figs. 2 and 3, it can be seen that the self-inductances do not change with respect to rotor position since the air gap between rotor



and stator is fixed. So, the rotation in the squirrel-cage IM and surface-mounted (SM) PMSM is due to the reaction torque.

When the rotor magnets are placed inside the rotor iron in PMSM, the machine is called interior permanent magnet (IPM). In this case, the magnetic reluctance is not uniform as it was for SM-PMSM. So, the self-inductance value of the stator will depend on the rotor position and, consequently, there will be nonzero reluctance torque in addition to the reaction torque. Therefore, an IPM has high efficiency and torque density because the motor can utilize both reaction and reluctance torque due to the magnetic saliency [2]. The interior permanent magnet motors have demonstrated attractive performance in the field-weakening region as well. This is accomplished by the virtue of the reluctance torque which allows for an extended speed range in the constant power region. Table 1 briefly compares PMSM and IM in terms of important characteristics including torque, cooling, fault tolerance, supply chain, speed ratio, and noise.

Equation 13 also suggests that the electromagnetic torque can be expressed symbolically as:

$$T = \sum_i f_i(I_s, I_r) g_i(\theta_r) \quad (16)$$

The above equation illustrates the impact of power electronic-based control (i.e.,  $f(\cdot, \cdot)$ ) and magnetic design (i.e.,  $g(\cdot)$ ). Power electronic-based control allows for injection of a desired waveform for electric current and, as such, many researchers have tried to obtain optimal waveforms which results in maximum torque, minimum torque pulsation, maximum torque per ampere, etc. For instance, in the case of PMSM, a maximum torque production has pointed toward the following relationship:

$$i_s \propto \frac{\partial \lambda_{pm}}{\partial \theta_r} \quad (17)$$

where  $\lambda_{pm}$  represents the flux linkage in the stator phase due to permanent magnets. Similarly, in switched reluctance drives, the stator current is found to be optimal when the following condition is met:

$$i_s^2 \propto \frac{\partial L_s}{\partial \theta_r} \quad (18)$$

where  $L_s$  represents the self-inductance of the stator phase. In general, the problem can be mathematically formulated as the following optimization task:

$$(I_s, I_r) = \text{Arg} \left( \text{Max} \left( \int_0^\pi \sum_i f_i(I_s, I_r) \cdot g_i\left(\frac{P}{2}\theta_r\right) d\theta_r \right) \right) \quad (19)$$

where  $P$  represents the number of magnetic poles. The boundary conditions imposed by the available DC-link voltage, time constants of the circuit, and power electronic circuits have to be taken account in this optimization process. Similar expressions can be found for other objective functions. It is clear that magnetic configuration of the electromechanical energy conversion device plays a central role in this process. In particular, it influences the room for optimization using power electronic converter.

## Permanent Magnet Motor Drives

### Classification

Permanent magnet synchronous machines (PMSM) also known as brushless DC machines (BLDC) form an important group of electric machinery that are being considered for variety of applications including electric propulsion of modern vehicles. PMSM is an electromechanical energy conversion device formed by

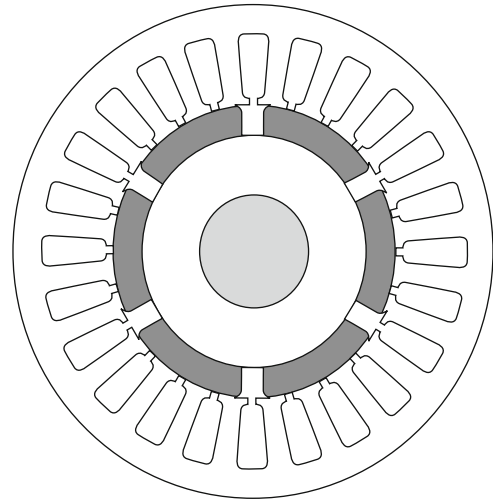
**AC Machines: Permanent Magnet Synchronous and Induction Machines. Table 1** Comparison of characteristics in PMSM and IPM

|         | [N-m/kg] | [W/kg] | Cooling | Fault tolerance | Supply chain | Speed ratio | Noise |
|---------|----------|--------|---------|-----------------|--------------|-------------|-------|
| IPM     | ++       | ++     | —       | —               | --           | +           | +     |
| SM-PMSM | ++       | +      | —       | —               | --           | —           | +     |
| IM      | +        | +      | --      | —               | ++           | +           | +     |

a conventional AC stator (i.e., sinusoidally distributed multiphase winding fed with a balanced set of currents). In this class of machines, permanent magnets are mounted on the rotor which then removes the need for injection of the current to a moving component. Special magnetic arrangement for the rotor provides the advantages of (a) brushless operation, (b) controllability of a DC machine, (c) high power density in constant torque region and also (d) compatibility with the commonly used multiphase inverters. The magnets can be placed in number of ways on the rotor. Depending on the placement, they are called either a surface-mounted PMSM (SM-PMSM) or an interior PMSM (IPM). In SM-PMSM, permanent magnets are placed on the rotor surface, making it easy to build, and skewed poles are easily magnetized on this surface-mounted type to minimize cogging torque.

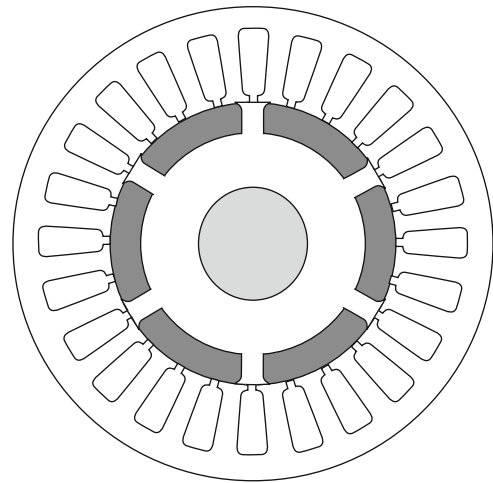
This configuration is used for low-speed applications because of the limitation that the magnets may fly apart if they are used in high-speed applications. Introduction of container shields to protect the permanent magnets will effectively increase the air gap length which undermines an efficient use of permanent magnets. These motors are considered to have small saliency, thus having practically equal inductances in both direct and quadrature axes. Besides, the relatively small self-inductance of SM-PMSM will result in a small electrical time constant which in turn requires fast switching for direct control of the current. Notably, in SM-PMSM, reluctance torque is almost nonexistent and reaction torque is the main source of motion. The permeability of the permanent magnet is almost that of the air, thus for calculation of self-inductances, the magnetic material should be treated as an extension of the air gap. In SM-PMSM, the rotor has an iron core that may be made of punched laminations for simplicity in manufacturing. Thin permanent magnets are mounted on the surface of this core using adhesives. Alternating magnets of the opposite magnetization direction produce radially directed flux density across the air gap. The magnetic field produced by these PMs then interacts with magnetic field produced by the stator current to produce torque. Figure 4 shows how magnets are placed in SM-PMSM [3].

Interior PMSMs have embedded permanent magnets which are mounted inside rotor as shown in Fig. 5 [3]. Unlike SM-PMSM, each PM is mounted inside the



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 4**

Cross-sectional area of an SM-PMSM



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 5**

Cross-sectional area of an IPM

rotor in IPM. This will make IPM a good candidate for high-speed applications. There is inductance variation for this type of rotor because the permanent magnet is equivalent to air in the magnetic circuit calculation. The use of axial lamination will further intensify the magnetic saliency for this family of machines. These machines are considered to have magnetic saliency in the sense that  $q$ -axis inductance is greater than the

$d$ -axis inductance. The  $q$  and  $d$  axes represent the quadrature and direct axes values, respectively. This frame of references rotate with the mechanical speed of the motor and as such a DC current in  $q$  and  $d$  axes can produce the same field as in the stationary AC windings of the machine. Since the PMs are not uniformly distributed on the rotor surface, the effective air gap between rotor and stator is not constant. As a result, the stator self-inductance varies with respect to rotor position. Consequently, in IPM, both reaction torque and reluctance torque contribute to the motion.

PMSM machines have recently attracted considerable attention from various sectors of industry due to their compact, efficient, and maintenance-free operation. Recent advancements in manufacturing of magnets have enabled the use of such machines in a wide range of applications. This in turn, has initiated many new opportunities for this family of electric machines.

These machines can further be classified into either sinusoidal type or trapezoidal type according to the type of flux produced in the air gap by the magnets. The difference is created by the way the magnets are magnetized initially before mounting on the rotor. Machines with sinusoidal back-emf can be controlled to achieve ripple-free torque whereas machines with nonsinusoidal back-emf offer lower inverter size and a reduction in losses.

### Permanent Magnet Material

Permanent magnet used in PMSM is a piece of magnetic material which, once magnetized or “charged” by an external magnetic field, retains a usefully large magnetic moment after the magnetizing force is removed. Thus, a permanent magnet becomes a source of magnetic field in the rotor which can interact with the magnetic field generated by stator currents to produce torque. For a magnet to be technologically useful, this magnetization must persist in the presence of fairly high-opposing fields (the permanent magnet must have a high “coercive force”) and often at elevated and sometimes at very cold temperatures. Its magnetic properties must be reasonably stable for long periods of time in adverse environments.

NdFeB magnets set a vivid example of this trend by offering an impressive energy density of about  $200\text{--}440 \frac{\text{kJ}}{\text{m}^3}$ . In addition, highly efficient performances,

more compact geometry, lower weight, and shorter electrical time constants are among other attractive attributes of these technologies.

Permanent magnets are described by their demagnetization characteristic. This is a portion of the hysteresis cycle located in the second quadrant of the flux density vs. field intensity (B-H) plane. Figure 6 shows demagnetization characteristics of various magnets at rated temperature ( $20^\circ\text{C}$ ) [5]. As can be observed, the demagnetization curve of a permanent magnet can be fully characterized by the following parameters:

1. Remanent magnetism ( $B_r$ ), where characteristic crosses the B-axis.  $B_r$  is the magnetic flux density remaining in a material, especially a ferromagnetic material, after removal of the magnetizing field.
2. Coercive field intensity ( $H_c$ ), where characteristic crosses the H-axis.  $H_c$  shows the material's resistance to becoming demagnetized.
3. Curvature connecting these two points, mostly one or two linear segments.

In fact, the second linear region that is close to the H-axis is considered as an unstable region.

Therefore a proper operating point will be located in the first linear part, which is expressed as:

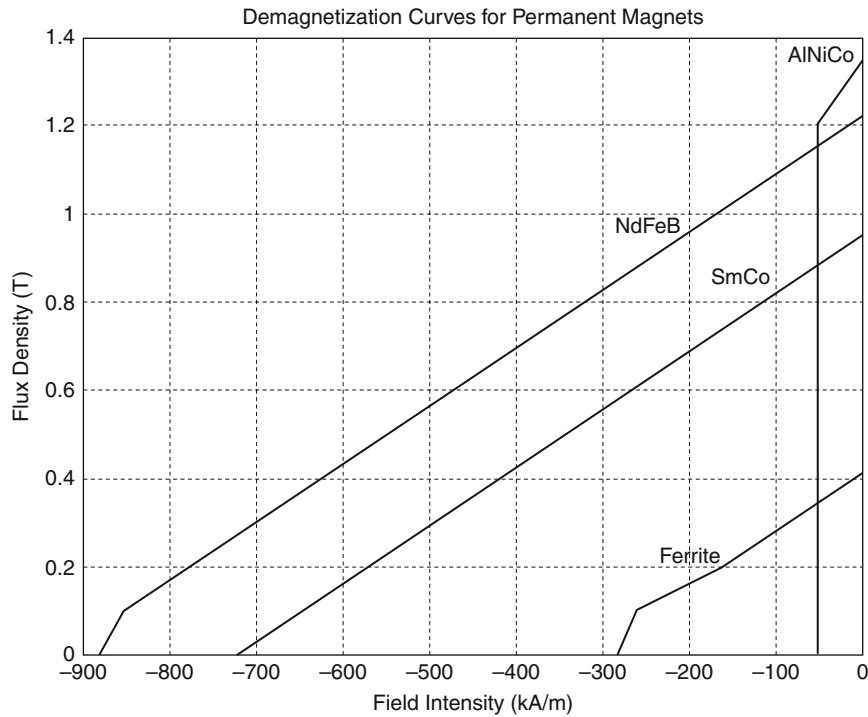
$$B = B_r + \mu_0\mu_r H \quad (20)$$

where  $\mu_r$  and  $\mu_0$  stand for relative and air permeability, respectively. It must be noted that in permanent magnets,  $\mu_r$  is very close to 1 (relative permeability of the air). Accordingly, the energy density of a permanent magnet (assuming a linear characteristic) can be computed as:

$$E = B.H \Rightarrow E_{Max} = \frac{B_r^2}{4\mu_0\mu_r} \left[ \frac{\text{kJ}}{\text{m}^3} \right]. \quad (21)$$

Figure 7 illustrates the constant energy loci along with demagnetization characteristics of various permanent magnets [5]. AlNiCo alloys, the hexaferrites, and the rare earth permanent magnets are the three types of magnets that are used in electrical machines among which rare earth PMs have attracted more attention in PMSM applications.

Chemically, the AlNiCos are iron–cobalt–nickel–based alloys with minor additions of aluminum and copper, and in some grades titanium. The ferrites have

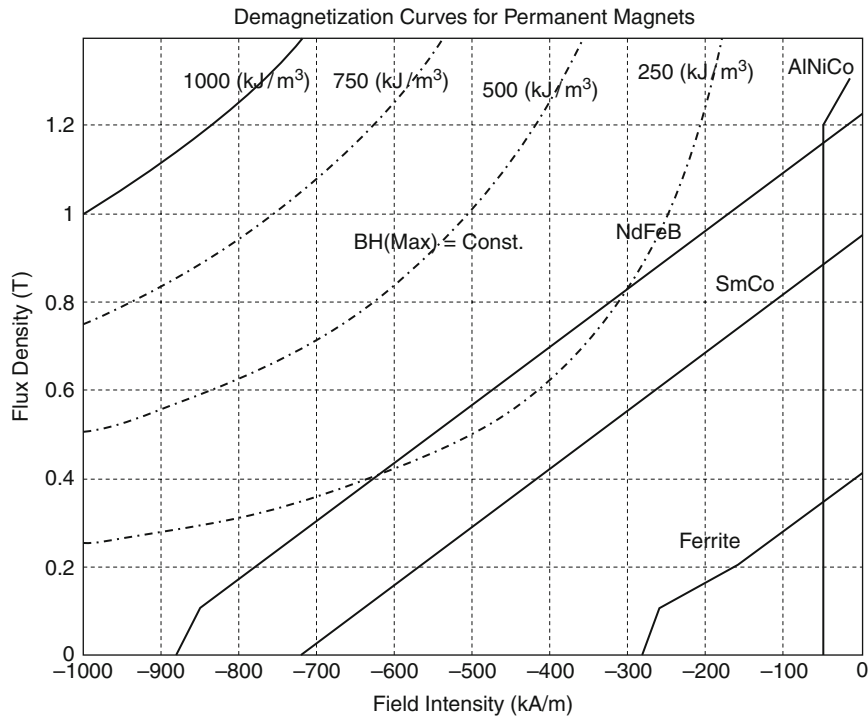


**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 6**  
Demagnetization characteristics for various permanent magnets

some peculiar properties setting them apart from other magnets, for better and worse. As oxides, they are electrical insulators, which eliminate eddy currents. (Most permanent magnet materials, being metals, are fairly good conductors.) This is advantageous under dynamic operating conditions as in electric motors and also facilitates pulsed field charging. They are chemically inert, so the production steps of milling, pressing, and surface grinding can be done in air and with water lubrication; they also have no corrosion problems in use even at high temperatures, but the magnetization in ferrites drops more rapidly in the presence of heating.

Developed in the 1970s and 1980s, rare earth magnets are the strongest type of permanent magnets made, substantially stronger than ferrite or AlNiCo magnets. The magnetic field typically produced by rare-earth magnets can be in excess of 1.4 T, whereas ferrite or ceramic magnets typically exhibit fields of 0.5–1 T. There are two types of rare-earth PMs: neodymium magnets ( $\text{Nd}_2\text{Fe}_{14}\text{B}$ ) and samarium–cobalt (SmCo magnets). The newest family of hard magnetic alloys, rare-earth PM, includes not only the

well-known samarium–cobalt and neodymium–iron–boron magnets but also many other related materials. All are based on intermetallic compounds formed when 3d-transition metals (TM), notably the strongly magnetic Co and Fe, are alloyed with rare-earth elements (RE) of the 4f-transition series, (i.e., the “lanthanides” Ce, Pr, Nd, Sm, Gd, Dy, and others,) or their close chemical relatives La and Y. Addition of still another element, such as B, C, Si, or Ti, often helps form ternary compounds [4]. Addressing different types of permanent magnet in the above graphs, it must be mentioned that ferrite magnets represent a low-cost solution while offering a limited flux density. AlNiCo magnets, which are more expensive as compared to ferrite magnets, demonstrate a very high remanent magnetism. This, however, is undermined by very limited coercive field intensity. The most expensive SmCo magnets represent quality in every aspect including high remanent magnetism, large coercive field intensity, and a fully linear demagnetization characteristic. They also are known as highly stable in the presence of high-temperature variation. Finally, the



AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 7

Constant energy loci and demagnetization curves of various PM

NdFeB magnets demonstrate very high-energy density at room temperature. Furthermore, the cost associated with NdFeB is much less than SmCo magnets. However, coercive field intensity in NdFeB magnets is highly sensitive to temperature changes. This, in turn, results in an inadequate performance at high temperatures.

In general, the main objectives in selecting a permanent magnet for motor drive application can be summarized as:

- High-energy density
- A linear demagnetization characteristic in the entire vicinity of the second quadrant (B-H plane)
- High stability with respect to temperature
- High specific resistance to mitigate eddy currents
- Durability against corrosion and demagnetization
- Low cost

Although achieving all these attributes in a single magnet is not possible, proper design can help in optimizing the performance of the drive in the context of the application.

Table 2 summarizes the magnetic characteristics of four families of the permanent magnets.

Comparing various properties of permanent magnets, the following observation can be made:

- A rise in temperature will reduce the remanent magnetism in all magnet types. On a percentage basis, ferrite magnets seem to be the most sensitive magnets, while SmCo demonstrates the least sensitivity.
- Coercive field intensity portrays different behavior for various materials. While an increase in temperature results in significant decrease of coercive field in NdFeB, an opposite response is seen in ferrite magnets. Overall, SmCo offers the least sensitivity to temperature.
- SmCo is the heaviest and the most expensive alternative among all candidates. It also presents one of the lowest specific resistances, which translates to high eddy-current losses.
- AlNiCo offers the highest remanent magnetism. This, however, is mainly undermined by a very



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Table 2** Summary of attributes for various permanent magnets

| Parameters at 20°C                                      | Ferrite                              | AlNiCo          | SmCo5           | NdFeB           |
|---------------------------------------------------------|--------------------------------------|-----------------|-----------------|-----------------|
| Remanent magnetism ( $B_r$ [T])                         | 0.38...0.42                          | 0.61...1.35     | 0.85...1.0      | 1.0...1.23      |
| Coercive field intensity ( $H_c$ [KA/m])                | 390...280                            | 59...50         | 1000...1200     | 1600...960      |
| Max. energy density ( $BH_{\max}$ [kJ/m <sup>3</sup> ]) | 28...34                              | 13...62         | 140...200       | 195...280       |
| Temperature coefficient ( $K_B$ [%/C])                  | −0.2... − 0.23                       | −0.02           | −0.04... − 0.05 | −0.11... − 0.13 |
| Temperature coefficient ( $K_H$ [%/C])                  | +0.4... + 0.22                       | +0.03... − 0.07 | −0.25           | −0.6... − 0.8   |
| Reversible permeability ( $\mu_{rev}$ )                 | 1.05                                 | < 5             | 1.05            | < 1.2           |
| Density [ $\rho$ /(kg/dm <sup>3</sup> )]                | 4.6...4.9                            | 6.7...7.3       | 8.1...8.3       | 7.3...7.4       |
| Specific electric resistance [ $\rho$ /μΩcm]            | 10 <sup>12</sup> ...10 <sup>16</sup> | 40...70         | 50...60         | 140             |
| Cost (%)                                                | 10...15                              | 40...60         | 600...800       | 200...300       |

limited coercive field intensity and extremely high conductivity.

- The reversible permeability in most cases is close to 1.

### Impact of Temperature on Performance of PMSM

Temperature variations are an inevitable attribute of many industrial applications such as automotive products. In fact, in some automotive applications, a temperature range of −40 – 150°C is reported. This, in turn, can alter the properties of permanent magnets used in PMSM. Figures 8 and 9 show demagnetization curves for NdFeB and ferrite magnets at various temperatures. As can be seen, a rise in temperature reduces the remanent magnetism in both cases. However, a severe reduction in coercive field in NdFeB poses a real limitation for armature currents at high temperatures. This will clearly undermine the practicality of using NdFeB magnets at very high temperatures. It must be noted that a modest range of temperature, below 40°C, can be appropriate for employing the NdFeB. The ferrite magnets, on the other hand, depict a reverse behavior at high temperatures; they tend to have larger coercive fields. This, however, at very low temperatures may cause a problem. According to Table 2, some rare-earth permanent magnets demonstrate a superior performance in terms of temperature stability. However, this comes with a significantly higher cost. Therefore, if a modest range of

temperatures can be accommodated, employment of NdFeB magnets is recommended.

In order to quantify the effects of temperature change, the linear (stable) part of demagnetization characteristics is expressed as follows:

$$B = f(H) \cong B_r + \mu_0 \mu_r H \quad (22)$$

$$B_r \approx B_{r0}(1 + k_B T/100) \quad (23)$$

$$H_K \approx H_{K0}(1 + k_H T/100) \quad (24)$$

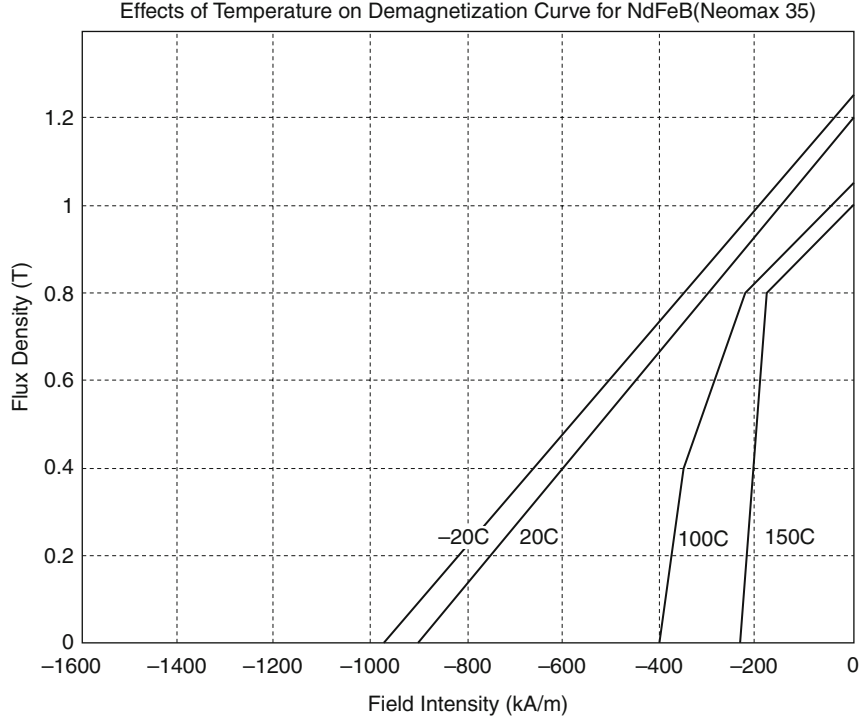
The no-load operating points are as follows:

$$B_{M0} \cong \frac{B_r}{1 + \left(\frac{\sigma}{h_M}\right)} = \frac{(B_{r0}(1 + k_B T/100))}{1 + \left(\frac{\sigma}{h_M}\right)} \quad (25)$$

in which  $\sigma$  and  $h_M$  represent air gap length and radial thickness of permanent magnets, respectively. A change in the temperature, therefore, directly shifts the location of operating point. This will affect the allowed armature current as the maximum coercive field has also been changed. This is expressed as:

$$\begin{aligned} \Delta H_M \leq H_{M0} - H_K &= \frac{-B_r}{\mu_0(1 + h_M/\sigma)} - H_K \\ &= \frac{-(B_{r0}(1 + k_B T/100))}{\mu_0(1 + h_M/\sigma)} - (H_{K0}(1 + k_H T/100)) \end{aligned} \quad (26)$$

One can say that an increase in temperature will reduce remanent magnetism and, hence, will reduce the torque productivity of the machine by a factor of:



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 8**  
Thermal behavior of NdFeB magnets

$$\frac{\partial B}{\partial T} = \left( \frac{k_B}{1 + (\sigma/h_M)} \right) \frac{B_{r0}}{100} \quad (27)$$

where  $T$  stands for temperature in degree Celsius. Table 2 provides a range of  $k_B$  values for various permanent magnets. At the same time, the permitted armature field to maintain a stable operation is limited by:

$$\frac{\partial B}{\partial T} = \left( \frac{-k_B B_{r0}}{\mu_0(1 + h_M/\sigma)} - k_H H_{K0} \right) \frac{1}{100} \quad (28)$$

It must be noted that a more strict limit on the armature excitation will further impact the torque generation of the permanent magnet synchronous motor drive. Table 2 provides a range of  $k_H$  values for various permanent magnets. In order to illustrate the effects of temperature, the following example is given. Table 3 shows three various designs (two-pole design is assumed) using ferrite, SmCo, and NdFeB permanent magnets.

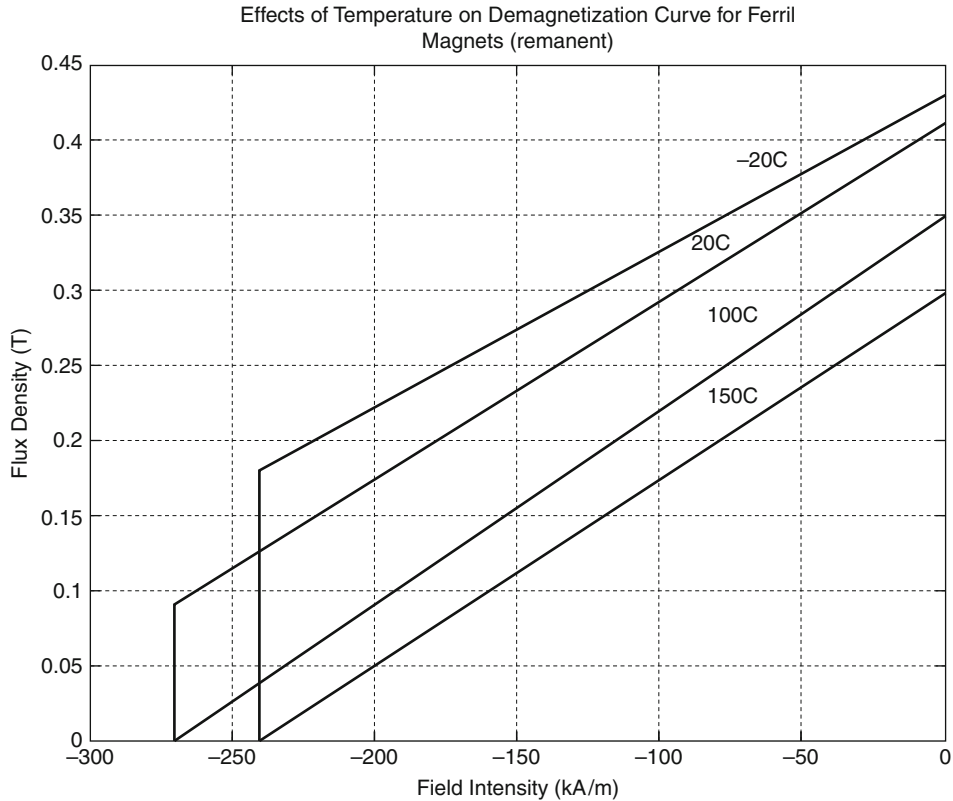
Once the air gap flux density and armature mmf are known, the following equation can be used to estimate electromagnetic torque:

$$T \propto \frac{\pi}{3} H B D^2 l \quad (29)$$

in which  $D$  and  $l$  represent outer diameter of the rotor and effective stack length of the machine, respectively. Using Eq. 29, one can select the stack length such that a total torque of 5 N - m is achieved. Table 4 shows the results of torque calculation.

In the next step, using the above method, operating points and the electromagnetic torque at  $-20$ ,  $20$ ,  $100$ , and  $150^\circ\text{C}$  are computed and listed in Table 5. Please note that the stack length of machines is kept constant in this exercise.

The results of this study are graphically shown in Fig. 10 [5]. As can be seen, reduction in remanent magnetism, due to temperature variations, is the main cause for reduction in torque. At high temperatures, despite the relatively low armature mmf, low coercive/knee point field intensity also contributes to



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 9**  
Thermal behavior of ferrite magnets

**AC Machines: Permanent Magnet Synchronous and Induction Machines. Table 3** Example design for illustration of temperature effects

| Type    | $\sigma$ (mm) | $h_M$ (mm) | D (mm) |
|---------|---------------|------------|--------|
| Ferrite | 0.5           | 4.5        | 50     |
| SmCo    | 0.5           | 1          | 50     |
| NdFeB   | 0.5           | 2          | 50     |

this reduction. It is therefore expected that a higher armature mmf would result in yet higher reduction in torque under this design. It must also be noted that SmCo depicts the most stable operation; however, with a reasonable range of temperatures, employment of NdFeB magnets seems to be the most appropriate choice. NdFeB magnets not only are better economic choices, they also offer higher specific resistance at a lower weight compared to SmCo magnets. This in

**AC Machines: Permanent Magnet Synchronous and Induction Machines. Table 4** Operational condition at 20°C

| Type    | $B_r$ (T) @ 20°C | $H_A$ (kA/m) @ 20°C | Stack length | Torque (N-m) |
|---------|------------------|---------------------|--------------|--------------|
| Ferrite | 0.42             | 30                  | 165          | 5            |
| SmCo    | 1                | 30                  | 96           | 5            |
| NdFeB   | 1.23             | 30                  | 64           | 5            |

turn minimizes the effects of high-frequency eddy-current losses caused by slot harmonics.

### Voltage and Torque Equations for SM-PMSM

Since SM-PMSM is more popular than IPM, this part is dedicated to describing dynamic equations of a typical two-pole three-phase SM-PMSM depicted in

**AC Machines: Permanent Magnet Synchronous and Induction Machines. Table 5** Effects of temperature on operational characteristics of PMSM

| Type    |     | $B_r$ (T) | $H_k$ (kA/m) | $B_{air\ gap}$ (T) | $H_M$ (kA/m) | Torque (N-m) |
|---------|-----|-----------|--------------|--------------------|--------------|--------------|
| Ferrite | -20 | 0.46      | -246         | 0.414              | -36          | 5.36         |
|         | 20  | 0.42      | -280         | 0.378              | -33          | 5            |
|         | 100 | 0.34      | -246         | 0.306              | -27          | 3.57         |
|         | 150 | 0.29      | -280         | 0.261              | -23          | 2.6          |
| SmCo    | -20 | 1.02      | -1,320       | 0.68               | -270         | 5.12         |
|         | 20  | 1         | -1,200       | 0.667              | -265         | 5            |
|         | 100 | 0.96      | -960         | 0.64               | -254         | 4.83         |
|         | 150 | 0.94      | -810         | 0.627              | -250         | 4.73         |
| NdFeB   | -20 | 1.29      | -980         | 1.032              | -205         | 5.18         |
|         | 20  | 1.23      | -900         | 0.984              | -196         | 5            |
|         | 100 | 1.1       | -220         | 0.88               | -175         | 4.42         |
|         | 150 | 1.02      | -190         | 0.816              | -162         | 3.82         |

Fig. 11 [6]. In this figure, stator includes three sinusoidally distributed windings placed  $120^\circ$  apart from each other. Each winding has  $N_s$  equivalent turns and resistance  $r_s$ . In this section, the voltage and torque equations that are used to describe the dynamic behavior of the PMSM will be established.

Voltage equations in machine variables are expressed through Eqs. 30–32

$$V_{as} = r_s i_{as} + p \lambda_{as} \quad (30)$$

$$V_{bs} = r_s i_{bs} + p \lambda_{bs} \quad (31)$$

$$V_{cs} = r_s i_{cs} + p \lambda_{cs} \quad (32)$$

Writing stator voltage equations in matrix format results in Eq. 33.

$$V_{abcs} = r_s i_{abcs} + p \lambda_{abcs} \quad (33)$$

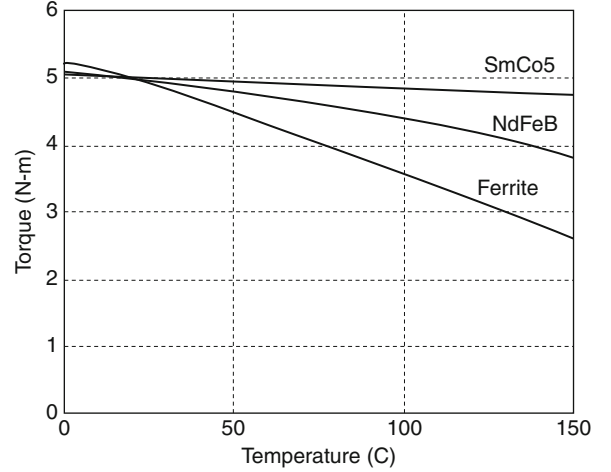
where

$$(f_{abcs})^T = [f_{as} \ f_{bs} \ f_{cs}] \quad (34)$$

$$r_s = \text{diag}[r_s \ r_s \ r_s] \quad (35)$$

The flux linkage can be written as:

$$\lambda_{abcs} = L_s i_{abcs} + \lambda'_m \quad (36)$$



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 10**

Effects of temperature on electromagnetic torque in brushed DC motor drives

where  $L_s$  is the stator inductance matrix and is expressed by Eq. 37. Also,  $\lambda'_m$  is the flux linkage established by the PMs on the rotor as viewed from stator windings and is expressed by Eq. 38. Where  $\lambda'_m$  is the flux linkage amplitude produced by the PMs as viewed from stator phase winding.

$$L_s = \begin{bmatrix} L_{aa} & L_{ab} & L_{ac} \\ L_{ba} & L_{bb} & L_{ab} \\ L_{ca} & L_{cb} & L_{cc} \end{bmatrix} \quad (37)$$

Where;

$$L_{ab} = L_{ba}$$

$$L_{ac} = L_{ca}$$

$$L_{bc} = L_{cb}$$

And;

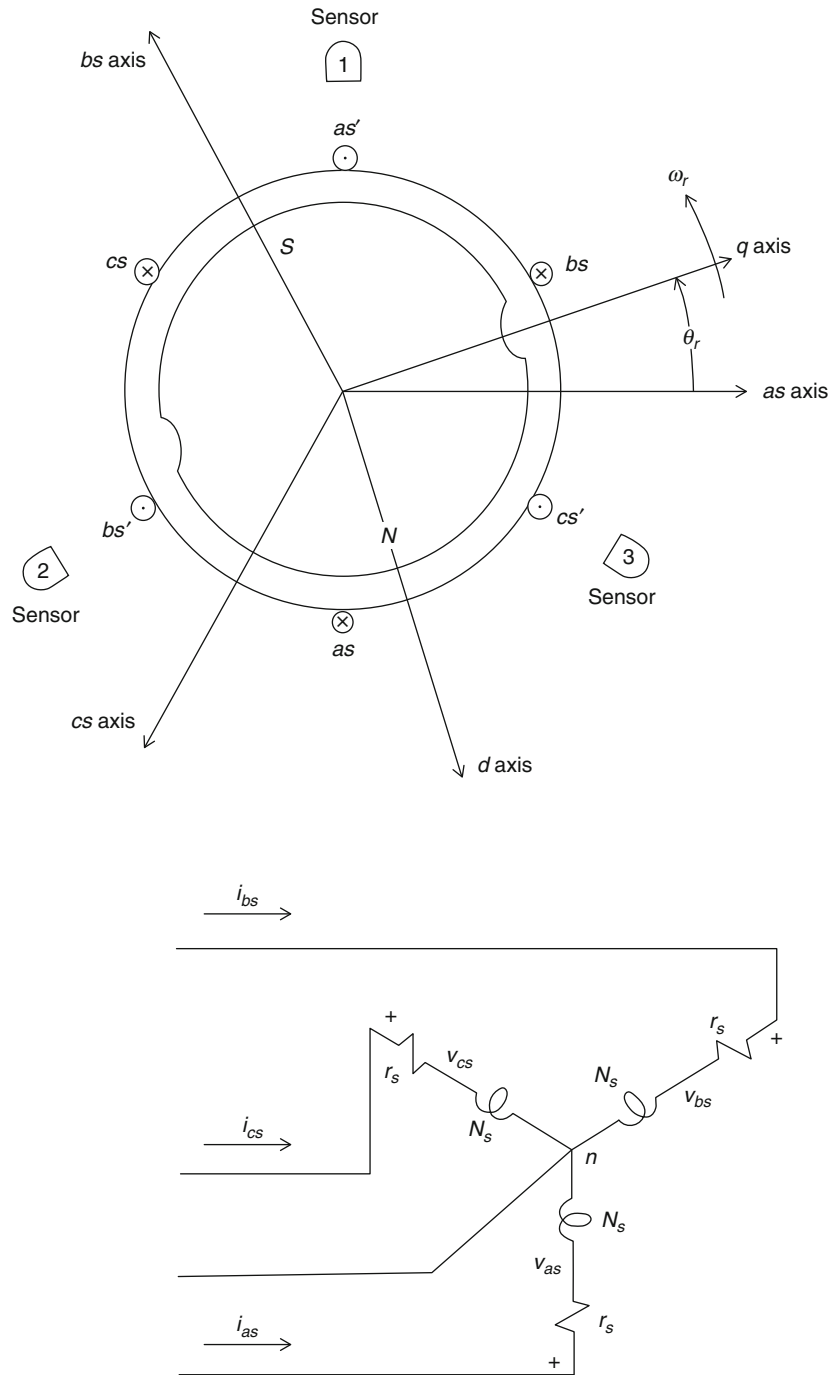
$$L_{aa} = L_{ls} + L_A - L_B \cos 2\theta_r$$

$$L_{ab} = -\frac{1}{2} L_A - L_B \cos 2\left(\theta_r - \frac{\pi}{3}\right)$$

$$L_{ac} = -\frac{1}{2} L_A - L_B \cos 2\left(\theta_r + \frac{\pi}{3}\right)$$

$$L_{bb} = L_{ls} + L_A - L_B \cos 2\left(\theta_r - \frac{2\pi}{3}\right)$$

$$L_{bc} = -\frac{1}{2} L_A - L_B \cos 2(\theta_r + \pi)$$



AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 11  
A two-pole three-phase SM-PMSM



$$L_{cc} = L_{ls} + L_A - L_B \cos 2\left(\theta_r + \frac{2\pi}{3}\right)$$

$$\lambda'_m = \lambda'_m \begin{bmatrix} \sin \theta_r \\ \sin(\theta_r - \frac{2\pi}{3}) \\ \sin(\theta_r + \frac{2\pi}{3}) \end{bmatrix} \quad (38)$$

So,  $p\lambda'_m$  would be the open-circuit voltage induced in each stator phase winding. In Eq. 38, it is assumed that the voltages induced by PMs on the stator windings are sinusoidal.

The electromagnetic torque in machine variables can be expressed by Eq. 39.

$$T_e = \left(\frac{P}{2}\right) \left\{ \frac{(L_{md} - L_{mq})}{3} \times \left[ \left( i_{as}^2 - \frac{1}{2} i_{bs}^2 - \frac{1}{2} i_{cs}^2 - i_{as} i_{bs} - i_{as} i_{cs} + 2 i_{bs} i_{cs} \right) \sin 2\theta_r \right. \right. \\ \left. \left. + \frac{\sqrt{3}}{2} (i_{bs}^2 i_{cs}^2 - 2 i_{as} i_{bs} + 2 i_{as} i_{cs}) \cos 2\theta_r \right] \right. \\ \left. + \lambda'_m \left[ \left( i_{as} - \frac{1}{2} i_{bs} - \frac{1}{2} i_{cs} \right) \cos \theta_r \right. \right. \\ \left. \left. + \frac{\sqrt{3}}{2} (i_{bs} - i_{cs}) \sin \theta_r \right] \right\} \quad (39)$$

The dynamic  $dq0$  modeling is used to study the motor during transient and steady state. This can be done by the virtue of Park's transformation. As discussed earlier, Park's transformation converts the three-phase voltage, current, and flux variables into  $dq0$  variables. Park's transformation for the three-phase stator voltages can be expressed as Eq. 40. Consequently, voltage equations in  $dq0$  frame are expressed as Eq. 41.

$$\begin{bmatrix} V_q \\ V_d \\ V_0 \end{bmatrix} = \frac{2}{3} \begin{bmatrix} \cos \theta_r & \cos(\theta_r - 120) & \cos(\theta_r + 120) \\ \sin \theta_r & \sin(\theta_r - 120) & \sin(\theta_r + 120) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} \quad (40)$$

$$V_{dq0s} = r_s i_{dq0s}^r + \omega_r \lambda_{dq0s}^r + p \lambda_{dq0s}^r \quad (41)$$

where

$$\left( \lambda_{dq0s}^r \right)^T = \begin{bmatrix} \lambda_{ds}^r & -\lambda_{qs}^r & 0 \end{bmatrix} \quad (42)$$

$$\lambda_{dq0s}^r = \begin{bmatrix} L_{ls} + L_{mq} & 0 & 0 \\ 0 & L_{ls} + L_{md} & 0 \\ 0 & 0 & L_{ls} \end{bmatrix} \begin{bmatrix} i_{qs}^r \\ i_{ds}^r \\ i_{0s}^r \end{bmatrix} + \lambda_m^r \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad (43)$$

Expanding these equations,

$$v_{qs}^r = r_s i_{qs}^r + \omega_r \lambda_{ds}^r + p \lambda_{qs}^r \quad (44)$$

$$v_{ds}^r = r_s i_{ds}^r - \omega_r \lambda_{qs}^r + p \lambda_{ds}^r \quad (45)$$

$$v_{0s} = r_s i_{0s} + p \lambda_{0s} \quad (46)$$

where

$$\lambda_{qs}^r = L_q i_{qs}^r \quad (47)$$

$$\lambda_{ds}^r = L_d i_{ds}^r + \lambda_m^r \quad (48)$$

$$\lambda_{0s} = L_{ls} i_{0s} \quad (49)$$

where  $L_q = L_{ls} + L_{mq}$  and  $L_d = L_{ls} + L_{md}$ .

Substituting Eq. 47 to Eq. 49 into Eq. 43 to Eq. 46, results in:

$$v_{qs}^r = (r_s + pL_q) i_{qs}^r + \omega_r L_d i_{ds}^r + \omega \lambda_m^r \quad (50)$$

$$v_{ds}^r = (r_s + pL_d) i_{ds}^r - \omega_r L_q i_{qs}^r \quad (51)$$

$$v_{0s} = (r_s + pL_{ls}) i_{0s} \quad (52)$$

The expression for electromagnetic torque in  $dq0$  reference frame is then written as Eq. 53).

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) (\lambda_{ds}^r i_{qs}^r - \lambda_{qs}^r i_{ds}^r) \quad (53)$$

Substituting Eq. 47 and Eq. 48 into Eq. 53, Eq. 54 is achieved as the final expression for the electromagnetic torque in  $dq0$  reference frame.

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) \left[ \lambda_m^r i_{qs}^r + (L_d - L_q) i_{qs}^r i_{ds}^r \right] \quad (54)$$

The second term in Eq. 54 denotes the reluctance torque which is not significant in SM-PMSM, and therefore, the equation for the electromagnetic torque in SM-PMSM can be simplified to:

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) \lambda_m^r i_{qs}^r \quad (55)$$

In which  $P$ ,  $\lambda_m^r$ , and  $i_{qs}^r$  represent number of magnetic poles, flux linkage due to permanent magnet as observed by the coils on the new frame of reference

(which is mounted on the rotor), and the quadrature component of the new two-phase system mounted on the rotor, respectively. This equation resembles that of a separately excited DC machine in which the armature current is replaced with. Although there is a difference between the separately excited brushed machine and SM-PMSM in the new frame of references and that is the fact illustrated in Eq. 50 and Eq. 51. As can be seen, the current in  $d$ -axis will influence the voltage equation in  $q$ -axis and visa verse. In separately excited brushed machine, the flux generated by the armature current does not link the field windings and as such does not appear in the phase voltage equation. From Eq. 55, it is evident that increasing the  $d$ -axis current does not contribute to the torque generation, and therefore, it is usually regulated around zero. However, to extend the constant power region of the machine, a negative value of the direct axis current can counteract the field of the permanent magnets leading to a smaller motional back-emf. This will allow for regulation of the current in stator phases with a penalty on efficiency. The second term in Eq. 54 represents the reluctance torque and only exists in interior permanent magnet machines. In this case, direct and quadrature axis current will both contribute to the formation of electromagnetic torque, and as such, they have to be optimally controlled to ensure a maximum torque per ampere operation. Due to the presence of the reluctance torque component, an extended operation in the constant power region is feasible. The latter makes IPM an attractive candidate for electric propulsion of electrified power trains. Figure 12 illustrates a graphical

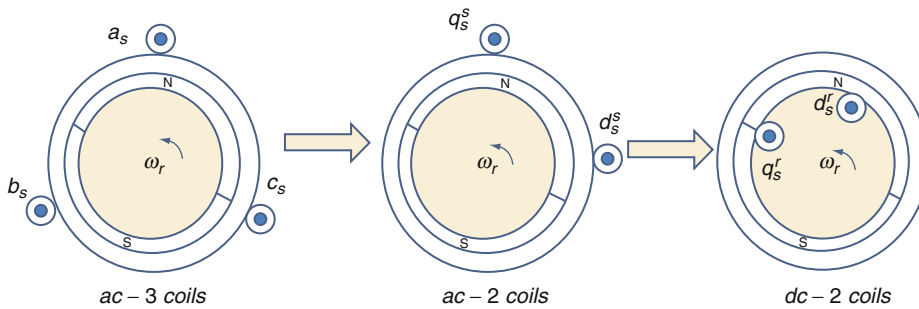
interpretation of change in frame of references in SM-PMSM.

### Field-Oriented Control in PMSM

Field-oriented control was introduced in the beginning of 1970s. It demonstrated that an induction motor or synchronous motor could be controlled similar to that of a separately excited DC motor by the orientation of the stator mmf or current vector in relation to the rotor flux to achieve a desired objective. The stator windings of the PMSM are fed by an inverter that generates a variable frequency and variable voltage magnitude. In field-oriented control of PMSM, instead of controlling the inverter frequency independently, the frequency and phase of the output wave are controlled using position sensors that are shown in Fig. 11 [4].

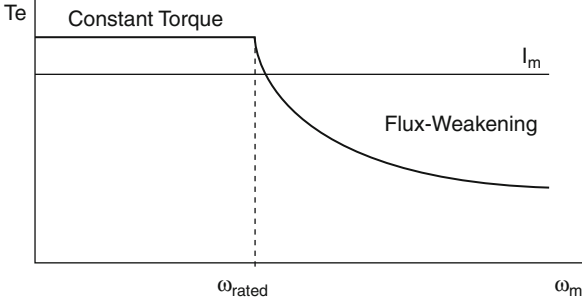
In order for the motor to behave like DC motor, the controller needs to know the position of the instantaneous rotor flux or rotor position of PMSM. This can be accomplished using an optical encoder. Having continuous access to the rotor position, the three-phase currents can be calculated.

There are generally two control modes for controlling PMSM. These control modes correspond to constant torque and flux weakening. These modes are determined by the physical limitation of the motor and the inverter. The limit is established by the rated speed of the motor, at which the constant torque operation finishes (i.e., back-emf partially exceeds the available inverter-phase voltage) and the flux weakening starts as shown in Fig. 13.



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 12**

Consecutive change in frame of references to reach a field-oriented arrangement with DC currents in quadrature and direct axes



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 13**

Torque versus speed characteristics of a PMSM

The field-oriented control of PMSM is also called vector control due to the fact that the direct and quadrature components of the current are controlled simultaneously. As shown in Fig. 14, accurate measurement of three-phase currents and rotor position is an integral part of this control strategy. Once the transformation to the synchronous frame of references (tied to the rotor) is completed, the desired values of the direct and quadrature current will be compared to the computed values of these DC quantities. The existing error will be then processed by a proportional + integral (i.e., PI) controller, and the controlled variable is then transformed back to the physical quantities (AC quantities) which are fed to the inverter control. According to Equation 55, the quadrature axis current of the stator will be controlled directly to achieve the desired torque in constant torque region. The direct axis current will be kept at zero to maximize the torque per ampere and to eliminate any need for reactive power [7].

In order to increase the speed range in PMSM, the flux in the  $d$ -axis current can be used to reduce the net flux generated by the permanent magnet. This action is called flux weakening.

For any given DC-link voltage, there exist a unique angular speed beyond which, at least for a nonzero portion of one electrical cycle, the induced back-emf (at rated constant torque) in each phase will exceed the available phase voltage. This speed is defined as base speed beyond which regulation of the current waveform will necessitate a net reduction in the back-emf. This can be accomplished by application of a negative direct axis current which will effectively reduce the flux originated from permanent magnet, thereby reducing

the motional back-emf. This method, however, introduces additional loss in the stator and cannot be practiced indefinitely, especially at high stator currents which may lead to partial demagnetization of the permanent magnets. The appearance of direct axis current will introduce the need for reactive power and a reduction in torque per ampere. This can be interpreted from the following equations:

$$\alpha = \tan^{-1} \left( \frac{i_q}{i_d} \right) \quad (56)$$

It is also known that:

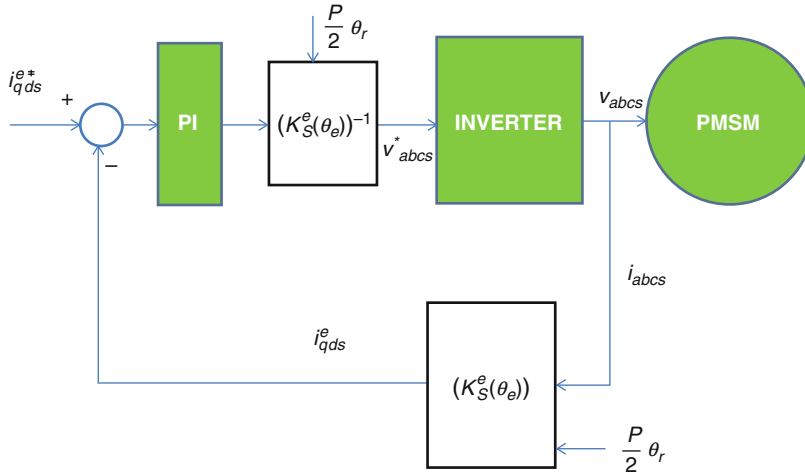
$$I_m = \sqrt{i_d^2 + i_q^2} \quad (57)$$

### AC Induction Machines

The induction machine is widely used in the industry for a variety of applications for converting electrical energy into mechanical work. Until the early 1970s, they had been mostly operated in the constant-voltage, uncontrolled mode which even today is the most common approach in practice.

The advent of thyristors and, consequentially, the development of variable-frequency converters allowed the wide-range speed control of induction motors. The scalar control method (e.g., Volts/Hertz control method) requires the simultaneous adjustments of the frequency and the magnitude of the sinusoidal voltage or current supplied to the motor. This allows making steady-state operating characteristics of an induction motor similar to those of a DC motor. Adjustable-speed AC drive systems, employing scalar control principles have been replacing DC drives in numerous industrial applications, such as pumps, fans, compressors, etc. Induction machines are extensively preferred as they are less expensive and more robust.

The scalar control of an induction motor does not make it dynamically equivalent to a DC motor. For instance, in an induction machine accurate position control is not possible since precise control of the instantaneous torque is needed to obtain the required speed trajectory. This, in turn, means that it is the instantaneous stator current which needs to be controlled. Scalar control applies to steady-state operation of the motor only, while no particular control strategies are involved for the transient operating conditions.

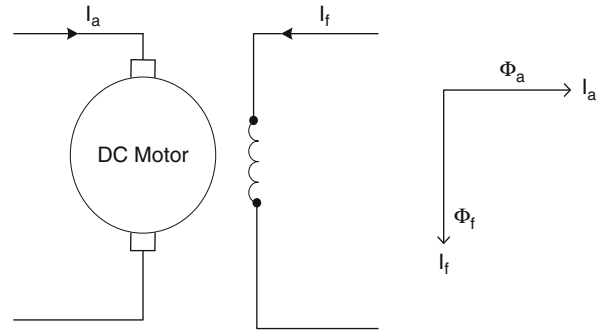


**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 14**  
Block diagram of the field-oriented (vector-controlled) control of SM-PMSM drive

The separately excited DC motor has physically separated armature and field windings whose mmfs are orthogonally displaced from each other in space. The armature current can be controlled without affecting the field current. Therefore, a decoupled control of torque and flux is possible, which results in fast torque and speed response. This is not the case in induction motors. To obtain the optimal torque from an induction motor, an algorithmic equivalent of the DC motor has to be established to obtain an appropriate decoupling between the stator and rotor flux. The control technique used to obtain such a stator–rotor physical link is named as field-oriented control or vector control.

The Field Orientation Principle was first formulated by Haase in 1968 and Blaschke in 1970. At that time, due to inadequate means to implement this mode, the concept of vector control seemed impractical. The advent of microprocessors made vector control of AC drives increasingly acceptable in the 1980s. Vector-controlled AC drive systems have proved to provide better dynamic performance than DC drive systems because of higher allowable speeds and shorter time constants of AC motors.

The typical induction motor is supplied by three alternating currents displaced by  $120^\circ$  which are then applied to a set of stationary coils. The resulting flux from the stator induces alternating currents in the “squirrel-cage” conductors of the rotor to create its



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 15**  
DC machine

own field. These fields interact to create torque. Thus the rotor currents in an AC induction motor cannot be directly controlled from an external source. As a result, the torque, which is dependent on the interaction of the stator field and the consequential rotor current, cannot be inherently controlled to obtain a desirable torque.

Vector control or field-oriented control of an AC induction motor is analogous to the control of a separately excited DC motor. In a DC motor, as shown in Fig. 15, the field flux  $\Phi_f$  is produced by the field current  $I_f$ .  $I_a$  is the armature current which produces the armature flux  $\Phi_a$ . The field flux  $\Phi_f$  and the armature flux  $\Phi_a$  are perpendicular to each other. Thus

the fields are decoupled and stationary with respect to each other. The armature current is controlled to control torque. The field flux is unaffected thus allowing a fast-transient response.

The purpose of field-oriented control is to recreate these orthogonal components in the AC machine so as to control the torque-producing current separately from the magnetic flux-producing current in order to achieve the responsiveness of a DC machine.

### Fundamental Equations of Induction Machine

A two-pole, three-phase wye-connected symmetrical induction machine is considered for the purpose, as shown in Fig. 16 [6]. The stator windings are identical, sinusoidally distributed displaced by  $120^\circ$  from each other with  $N_s$  equivalent turns and resistance  $r_s$ . The rotor windings are also considered to be three identical sinusoidally distributed displaced by  $120^\circ$  from each other with  $N_r$  equivalent turns and resistance  $r_r$ .

The voltage equations for the stator and rotor circuits can be written as:

$$V_{as} = r_s i_{as} + p \lambda_{as} \quad (58)$$

$$V_{bs} = r_s i_{bs} + p \lambda_{bs} \quad (59)$$

$$V_{cs} = r_s i_{cs} + p \lambda_{cs} \quad (60)$$

$V_{as}$ ,  $V_{bs}$ , and  $V_{cs}$  are the three-phase balanced voltages which rotate at supply frequency.  $\lambda_{as}$ ,  $\lambda_{bs}$ , and  $\lambda_{cs}$  are the stator flux linkages.

$$V_{ar} = r_r i_{ar} + p \lambda_{ar} \quad (61)$$

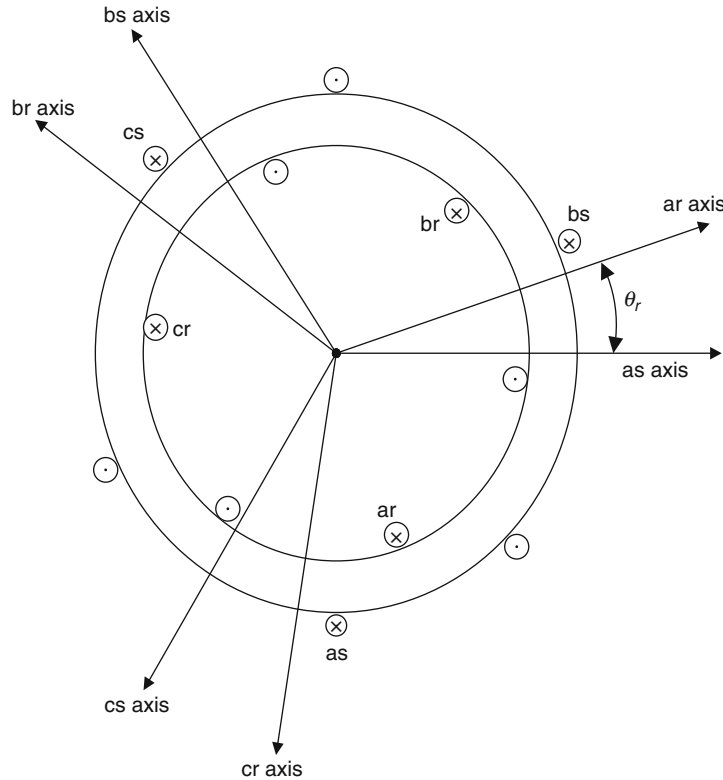
$$V_{br} = r_r i_{br} + p \lambda_{br} \quad (62)$$

$$V_{cr} = r_r i_{cr} + p \lambda_{cr} \quad (63)$$

$V_{ar}$ ,  $V_{br}$ , and  $V_{cr}$  are the rotor voltages and the rotor flux linkages which rotate at rotor speed,  $\omega_r$ .

The above equations can be written in short as:

$$V_{abcs} = r_s i_{abcs} + p \lambda_{abcs} \quad (64)$$



AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 16

Two-pole, three-phase symmetrical induction machine

$$V_{abcr} = r_s i_{abcr} + p \lambda_{abcr} \quad (65)$$

where

$$(V_{abcs})^T = [V_{as} \ V_{bs} \ V_{cs}] \quad (66)$$

$$(V_{abcr})^T = [V_{ar} \ V_{br} \ V_{cr}] \quad (67)$$

The subscripts “s” and “r” denote that the variables and parameters are associated with the stator and the rotor, respectively. The flux linkages may be written as:

$$\begin{bmatrix} \lambda_{abcs} \\ \lambda_{abcr} \end{bmatrix} = \begin{bmatrix} L_s & L_{sr} \\ (L_{sr})^T & L_r \end{bmatrix} \begin{bmatrix} i_{abcs} \\ i_{abcr} \end{bmatrix} \quad (68)$$

$L_s$  and  $L_r$  are the winding inductance matrices of the stator and the rotor, respectively.  $L_{sr}$  is the mutual inductance of the stator and rotor windings. The inductance matrix for the stator can be written as following:

$$L_s = \begin{bmatrix} L_{ls} + L_{ms} - 1/2L_{ms} & -1/2L_{ms} \\ -1/2L_{ms} & L_{ls} + L_{ms} - 1/2L_{ms} \\ -1/2L_{ms} & -1/2L_{ms} & L_{ls} + L_{ms} \end{bmatrix} \quad (69)$$

$L_{ls}$  and  $L_{ms}$  are the leakage and magnetizing inductances of the stator winding, respectively.

Similarly, the inductance matrix of the rotor winding is written as:

$$L_r = \begin{bmatrix} L_{lr} + L_{mr} - 1/2L_{mr} & -1/2L_{mr} \\ -1/2L_{mr} & L_{lr} + L_{mr} - 1/2L_{mr} \\ -1/2L_{mr} & -1/2L_{mr} & L_{lr} + L_{mr} \end{bmatrix} \quad (70)$$

$L_{lr}$  and  $L_{mr}$  are the leakage and magnetizing inductances of the rotor winding, respectively.

The mutual inductance matrix can be written as:

$$L_{sr} = L_{sr} \begin{bmatrix} \cos(\theta_r) & \cos(\theta_r + \frac{2\pi}{3}) & \cos(\theta_r - \frac{2\pi}{3}) \\ \cos(\theta_r - \frac{2\pi}{3}) & \cos(\theta_r) & \cos(\theta_r + \frac{2\pi}{3}) \\ \cos(\theta_r + \frac{2\pi}{3}) & \cos(\theta_r - \frac{2\pi}{3}) & \cos(\theta_r) \end{bmatrix} \quad (71)$$

$L_{sr}$  is the amplitude of the mutual inductance between the stator and rotor windings.  $\theta_r$  is the rotor position. For a machine with  $P$  number of poles,

$$\theta_r = \frac{P}{2} \theta_{rm} \quad (72)$$

where,  $\theta_{rm}$  is the mechanical rotor position. The speed  $\omega_r$  in terms of the rotor mechanical speed  $\omega_{rm}$  for a  $P$ -pole induction machine may therefore be expressed as:

$$\omega_r = \frac{P}{2} \omega_{rm} \quad (73)$$

It is convenient to express the physical variables of the rotor referred to the stator variables by appropriate turns-ratios:

$$V'_{abcr} = \frac{N_s}{N_r} V_{abcr} \quad (74)$$

$$i'_{abcr} = \frac{N_s}{N_r} i_{abcr} \quad (75)$$

$$\lambda'_{abcr} = \frac{N_s}{N_r} \lambda_{abcr} \quad (76)$$

The resistive and inductive elements of the rotor coil as referred to the stator windings may be written as:

$$L_{ms} = \frac{N_s}{N_r} L_{sr} \quad (77)$$

$$L'_{sr} = \frac{N_s}{N_r} L_{sr} \quad (78)$$

Thus, one can conclude that:

$$L'_{sr} = L_{ms} \begin{bmatrix} \cos(\theta_r) & \cos(\theta_r + \frac{2\pi}{3}) & \cos(\theta_r - \frac{2\pi}{3}) \\ \cos(\theta_r - \frac{2\pi}{3}) & \cos(\theta_r) & \cos(\theta_r + \frac{2\pi}{3}) \\ \cos(\theta_r + \frac{2\pi}{3}) & \cos(\theta_r - \frac{2\pi}{3}) & \cos(\theta_r) \end{bmatrix} \quad (79)$$

$$r'_r = \left(\frac{N_s}{N_r}\right)^2 r_r \quad (80)$$

$$L_{mr} = \left(\frac{N_s}{N_r}\right)^2 L_{ms} \quad (81)$$

$$L'_r = \left(\frac{N_s}{N_r}\right)^2 L_r \quad (82)$$

One may therefore write:

$$L'_r = \begin{bmatrix} L'_{lr} + L_{ms} - 1/2L_{ms} - 1/2L_{ms} \\ -1/2L_{ms} & L'_{lr} + L_{ms} - 1/2L_{ms} \\ -1/2L_{ms} - 1/2L_{ms} & L'_{lr} + L_{ms} \end{bmatrix} \quad (83)$$



where

$$L'_{lr} = \left( \frac{N_s}{N_r} \right)^2 L_{lr} \quad (84)$$

Thus the flux linkages may be written as:

$$\begin{bmatrix} \lambda_{abcs} \\ \lambda'_{abcr} \end{bmatrix} = \begin{bmatrix} L_{s'} & L'_{sr} \\ (L_{s'})^T & L_r \end{bmatrix} \begin{bmatrix} i_{abcs} \\ i'_{abcr} \end{bmatrix} \quad (85)$$

### Power Electronic Inverter

The inverter consists of six switching devices which produce three-phase AC power from a DC source. Figure 17 displays the inverter topology.

The DC voltage is assumed to be constant, and the transistors are assumed to be ideal switches. From Fig. 17 and Kirchhoff's Voltage Law,

$$V_{ag} = V_{as} + V_{ng} \quad (86)$$

$$V_{bg} = V_{bs} + V_{ng} \quad (87)$$

$$V_{cg} = V_{cs} + V_{ng} \quad (88)$$

Adding equations Eq. 86 to Eq. 88 yields,

$$V_{ag} + V_{bg} + V_{cg} = V_{as} + V_{bs} + V_{cs} + 3V_{ng} \quad (89)$$

By Kirchhoff's Current Law,

$$i_{as} + i_{bs} + i_{cs} = 0 \quad (90)$$

which implies,

$$\frac{di_{as}}{dt} + \frac{di_{bs}}{dt} + \frac{di_{cs}}{dt} = 0 \quad (91)$$

Thus, adding Eqs. 58 to 60 and substituting Eqs. 90 and 91 into this sum results in:

$$V_{as} + V_{bs} + V_{cs} = 0 \quad (92)$$

Thus Eq. 89 reduces to:

$$V_{ag} + V_{bg} + V_{cg} = 3V_{ng} \quad (93)$$

Solving Eq. 93 for  $V_{cg}$  and substituting this result into Eqs. 86–88 yields:

$$V_{as} = \frac{2}{3} V_{ag} - \frac{1}{3} V_{bg} - \frac{1}{3} V_{cg} \quad (94)$$

$$V_{bs} = \frac{2}{3} V_{bg} - \frac{1}{3} V_{ag} - \frac{1}{3} V_{cg} \quad (95)$$

$$V_{cs} = \frac{2}{3} V_{cg} - \frac{1}{3} V_{ag} - \frac{1}{3} V_{bg} \quad (96)$$

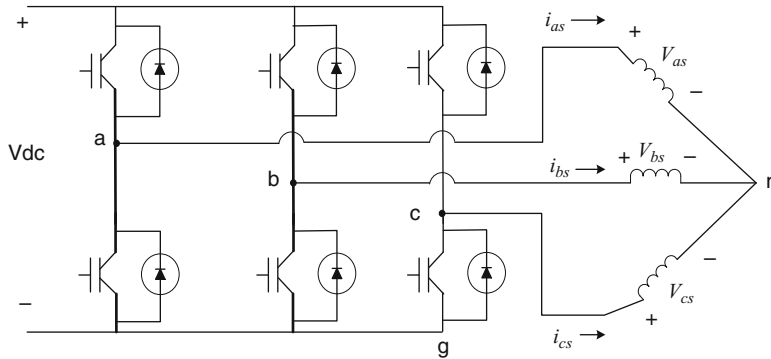
Thus, the given inverter line-to-ground voltages, the machine line-to-neutral voltages may be calculated using Eqs. 94–96.

### IM Drive in Synchronous Reference Frame

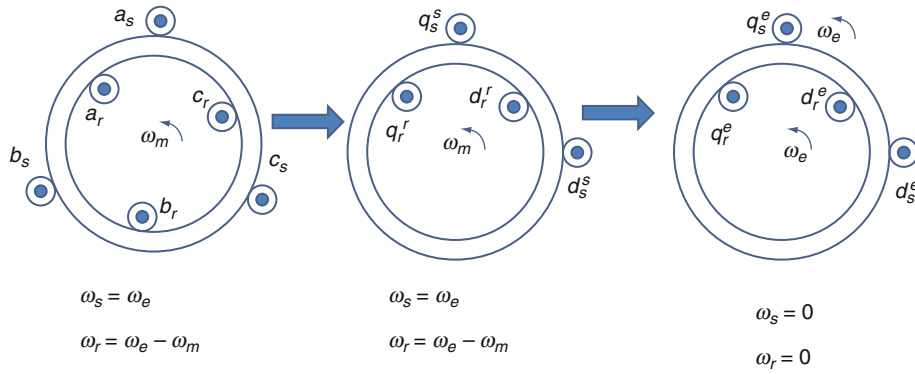
To reduce the complexity involved in analysis with these differential equations, one can apply the reference-frame theory to transform the variables and equations associated with the stator and the rotor to the synchronous reference frame. In the case of a synchronous reference frame, the reference frame rotates with the same frequency as the supply frequency  $\omega_e$  (see Fig. 18).

The transformation of variables to the synchronous reference frame can be obtained by using a transformation matrix  $K_s(\theta_e)$  such that:

$$f_{qd0} = K_s(\theta_e) f_{abc} \quad (97)$$



AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 17  
Inverter drive



**AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 18**  
Change in frame of references applied to three-phase induction machine

where  $f$  can be any of the variables  $V$ ,  $i$ , or  $\lambda$  and  $K_s(\theta_e)$  can be expressed as:

$$K_s(\theta_e) = \frac{2}{3} \begin{bmatrix} \cos(\theta_e) & \cos(\theta_e - \frac{2\pi}{3}) & \cos(\theta_e + \frac{2\pi}{3}) \\ \sin(\theta_e) & \sin(\theta_e - \frac{2\pi}{3}) & \sin(\theta_e + \frac{2\pi}{3}) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad (98)$$

where  $\theta_e = \int_0^t \omega_e(\xi) d\xi + \theta_{e0}$  and  $\theta_{e0}$  is the initial angle of the reference frame.

Applying this transformation to the stator voltage equations, one gets a set of equations transformed into the synchronous frame of reference which can be expressed as follows:

$$V_{qs} = r_s i_{qs} + p\lambda_{qs} + \omega_e \lambda_{ds} \quad (99)$$

$$V_{ds} = r_s i_{ds} + p\lambda_{ds} - \omega_e \lambda_{qs} \quad (100)$$

$$V_{os} = r_s i_{os} + p\lambda_{os} \quad (101)$$

Similarly, applying the transform to the rotor voltage equations, one obtains a set of equations transformed into the synchronous frame of reference which can be expressed as follows:

$$V'_{qr} = r'_r i'_{qr} + p\lambda'_{qr} + (\omega_e - \omega_r) \lambda'_{dr} \quad (102)$$

$$V'_{dr} = r'_r i'_{dr} + p\lambda'_{dr} + (\omega_e - \omega_r) \lambda'_{qr} \quad (103)$$

$$V'_{or} = r'_r i'_{or} + p\lambda'_{or} \quad (104)$$

In Eq. 99 to Eq. 101,  $V_{qs}$ ,  $i_{qs}$ , and  $\lambda_{qs}$  are  $q$ -axis stator variables.  $V_{ds}$ ,  $i_{ds}$ , and  $\lambda_{ds}$  are the  $d$ -axis stator variables

and  $V_{os}$ ,  $i_{os}$ , and  $\lambda_{os}$  are the zero-axis stator variables. Under balanced normal operating conditions, the zero-axis variables have zero values. Similarly, the variables with subscript “r” are the corresponding rotor variables.

The flux linkage equation after transformation may be expressed in the  $qdo$ -axis as:

$$\begin{bmatrix} \lambda_{qdos} \\ \lambda_{abcr} \end{bmatrix} = \begin{bmatrix} K_s L_s K_s^{-1} & K_s L_{sr} K_s^{-1} \\ K_s (L_{sr})^T K_s^{-1} & K_s L_r K_s^{-1} \end{bmatrix} \begin{bmatrix} i_{qdos} \\ i_{qdor} \end{bmatrix} \quad (105)$$

The terms of the matrix in Eq. 105 can be evaluated to obtain the following forms:

$$K_s L_s K_s^{-1} = \begin{bmatrix} L_{ls} + L_M & 0 & 0 \\ 0 & L_{ls} + L_M & 0 \\ 0 & 0 & L_{ls} \end{bmatrix} \quad (106)$$

where

$$L_M = \frac{3}{2} L_{ms} \quad (107)$$

$$K_s L_r K_s^{-1} = \begin{bmatrix} L'_{lr} + L_M & 0 & 0 \\ 0 & L'_{lr} + L_M & 0 \\ 0 & 0 & L'_{lr} \end{bmatrix} \quad (108)$$

$$K_s L_{sr} K_s^{-1} = \begin{bmatrix} L_M & 0 & 0 \\ 0 & L_M & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (109)$$

Thus the flux linkage equations in the synchronous frame are expressed as:

$$\lambda_{qs} = L_{ls} i_{qs} + L_M (i_{qs} + i'_{qr}) \quad (110)$$

$$\lambda_{ds} = L_{ls} i_{ds} + L_M (i_{ds} + i'_{dr}) \quad (111)$$

$$\lambda_{os} = L_{ls} i_{os} \quad (112)$$

$$\lambda'_{qr} = L'_{lr} i'_{qr} + L_M (i_{qs} + i'_{qr}) \quad (113)$$

$$\lambda'_{dr} = L'_{lr} i'_{dr} + L_M (i_{ds} + i'_{dr}) \quad (114)$$

$$\lambda'_{or} = L'_{lr} i'_{or} \quad (115)$$

The expression for the electromagnetic torque in terms of the reference frame variables can be written as:

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) L_M (i_{qs} i'_{dr} - i_{ds} i'_{qr}) \quad (116)$$

The electromagnetic torque in terms of flux linkages can be written as:

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) (\lambda'_{qr} i'_{dr} - \lambda'_{dr} i'_{qr}) \quad (117)$$

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) (\lambda_{ds} i_{qs} - \lambda_{qs} i_{ds}) \quad (118)$$

### Fundamental Equations of the IM Drive Under Field-Oriented Control

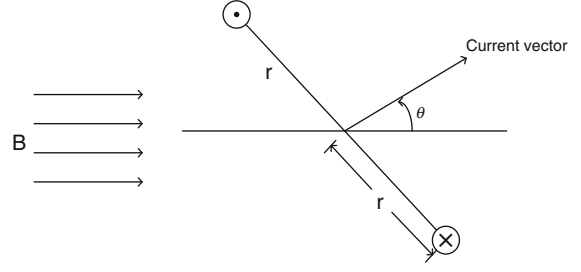
To understand the concept of vector control in an induction machine, a current loop consisting of a coil of length  $L$ , radius  $r$ , and with  $N$  number of turns carrying a current  $i$  in the presence of a flux density  $B$ , as shown in Fig. 19, is considered.

According to the Lorenz force equation, the torque acting on the current loop is given by:

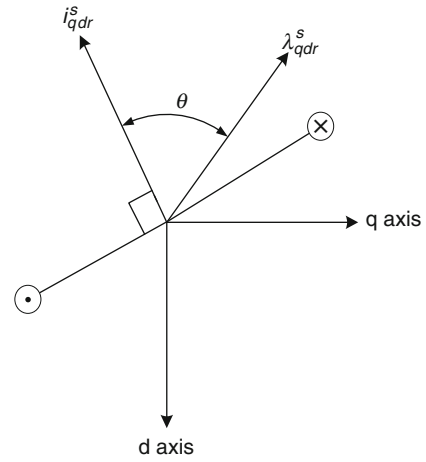
$$T_e = -2BiNLr \sin \theta \quad (119)$$

Thus, the magnitude of the torque is maximum when the current vector is perpendicular to the direction of flux density,  $B$ . The same concept may be extended to an induction machine. Consider Fig. 20 representing an induction machine torque production phenomenon. The rotor current and flux linkage vectors respectively are:

$$i'_{qdr} = [i'_{qr} \quad i'_{dr}]^T \quad (120)$$



AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 19  
Current loop



AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 20  
Induction machine torque production phenomenon

and

$$\lambda'_{qdr} = [\lambda'_{qr} \quad \lambda'_{dr}]^T \quad (121)$$

It is also known from Eq. 117

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) (\lambda'_{qr} i'_{dr} - \lambda'_{dr} i'_{qr}) \quad (122)$$

this may be expressed as:

$$T_e = \left(\frac{3}{2}\right) \left(\frac{P}{2}\right) |\lambda'_{qdr}| |i'_{qdr}| \sin \theta \quad (123)$$

This is analogous to Eq. 117. Thus, one can conclude that for a given magnitude of flux linkage, torque is maximum in the case when the rotor current vectors

are perpendicular to the flux linkage vectors. In vector control, the method to achieve the condition that the rotor flux and rotor current vectors are always perpendicular is twofold. The first condition is that:

$$\dot{\lambda}'_{qr} = 0 \quad (124)$$

This condition can be satisfied by choosing a suitable  $\theta_e$  on an instantaneous basis such that all the rotor flux is in the  $d$ -axis. The second condition is that:

$$\dot{i}'_{dr} = 0 \quad (125)$$

This condition can be achieved if the  $d$ -axis stator current is held constant.

Combining the  $d$ -axis stator and rotor flux linkage equations and the condition  $\dot{i}'_{dr} = 0$  results in:

$$\lambda'_{ds} = L_{ss} i'_{ds} \quad (126)$$

where

$$L_{ss} = L_{ls} + L_M \quad (127)$$

and

$$\dot{\lambda}'_{dr} = L_M \dot{i}'_{ds}. \quad (128)$$

Combining Eqs. 107 and 114, the torque equation may now be expressed as:

$$T_e = -\left(\frac{3}{2}\right) \left(\frac{P}{2}\right) (\dot{\lambda}'_{dr} i'_{qr}) \quad (129)$$

Since  $\dot{\lambda}'_{qr} = 0$ , from the  $q$ -axis rotor flux linkage equation, it is known that:

$$\dot{i}'_{qr} = -\frac{L_M}{L_{rr}} \dot{i}'_{qs} \quad (130)$$

where

$$L_{rr} = L_{lr} + L_M \quad (131)$$

Combining Eqs. 129 and 130 results in Eq. 132.

$$T_e = -\left(\frac{3}{2}\right) \left(\frac{P}{2}\right) \left(\frac{L_M}{L_{rr}}\right) (\dot{\lambda}'_{dr} i'_{qs}) \quad (132)$$

Equations 126–132 are the equations of the IM drive under rotor flux-oriented control.

## Field-Oriented Control Implementations

The basis of the field orientation algorithm is to use the flux angle, usually the rotor flux angle, to decouple the torque and flux components of the stator current. The most challenging feature of field orientation is the estimation or measurement of the flux angle. The direct field-oriented method, in which direct measurement of flux is performed using flux sensing coils and Hall-effect devices, proved to be impractical for general use. The indirect field-oriented control (IFOC) method has become more common. In this case, the flux angle is not measured directly but is estimated from the equivalent circuit model and from measurements of rotor speed, stator current, and the voltage. Given sufficiently correct current and position sensors, this method performs reasonably well over the entire speed range.

### Direct Field-Oriented Control

In direct field-oriented control, the magnitude and angular position of the reference flux vector are either measured or estimated from the stator voltage and currents using flux observers. Hall sensors can be used to measure magnetic fields. If the position of the stationary reference frame is set to zero then the rotor flux linkage in the  $q$ - $d$ -axes of the synchronous reference frame may be written as:

$$\begin{bmatrix} \lambda'_{qr} \\ \lambda'_{dr} \end{bmatrix} = \begin{bmatrix} \cos \theta_e & -\sin \theta_e \\ \sin \theta_e & \cos \theta_e \end{bmatrix} \begin{bmatrix} \lambda'^s_{qr} \\ \lambda'^s_{dr} \end{bmatrix} \quad (133)$$

In order to achieve  $\dot{\lambda}'_{qr} = 0$ , the position of the synchronous reference frame can be defined as:

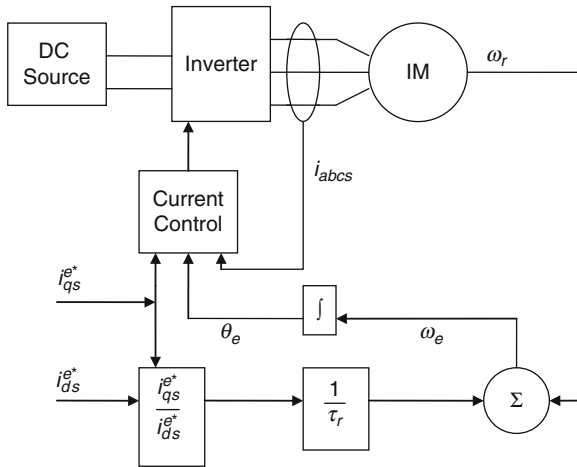
$$\theta_e = \text{angle}(\lambda'^s_{qr} - j\lambda'^s_{dr}) + \frac{\pi}{2} \quad (134)$$

$$\dot{\lambda}'_{dr} = \sqrt{(\dot{\lambda}'^s_{qr})^2 + (\dot{\lambda}'^s_{dr})^2} \quad (135)$$

The mutual flux linkages  $\lambda'^s_{qm}$  and  $\lambda'^s_{dm}$  may be regarded as measurable. The mutual flux linkage in the  $q$ -axis may be expressed as:

$$\lambda'^s_{qm} = L_M (\dot{i}'^s_{qs} + \dot{i}'^s_{qr}) \quad (136)$$





AC Machines: Permanent Magnet Synchronous and Induction Machines. Figure 22  
Indirect field-oriented control

## Future Directions

Induction motors are subject to many kinds of different internal and external faults during their operation [8]. Often the motor can run with these faults but ultimately it results in performance degradation of the machine, production shutdowns of the industrial process and sometimes permanent burnouts of the motor. Such shutdowns are costly in terms of lost production time, maintenance costs, and wasted raw materials. Therefore, schemes and methods for the diagnosis and detection of the external faults of IMs are major areas of research. Stator faults are one of the most commonly reported faults. Almost 30–40% of all reported faults are related to stator [9]. Stator faults are essentially insulation failures, which ultimately lead to commonly known line-to-ground or line-to-line faults. The primary causes for stator faults are high stator core or winding temperatures; slack core lamination; slot wedges and joints; loose bracing for end winding; contamination due to oil, moisture, and dirt; short-circuit or starting stresses; electrical discharges; and leakage in cooling systems, to name a few. Still, quite a few techniques are available to detect stator faults. So, detection of stator faults in induction machines has attracted much attention in recent years.

Fault-tolerant control for induction machine drives is another interesting issue since it enhances the reliability of the drive without inducing high-cost

investment. Improving reliability of electric motor drives can be achieved by improving the fault tolerance of converter switch modules and the fault tolerance of the motor control algorithm. Field-oriented fault-tolerant control algorithms and PWM strategies under fault conditions have been recently introduced, and their performances are studied [10].

Accurate assessment of the electro-magneto-mechanical energy conversion process is another significant field of research which usually requires a Multiphysics Finite Element (FEA) analysis [11]. Since FE analysis is a time-consuming numerical process, lumped parameter models, simplified equivalent magnetic circuits, and separated field analysis are usually used for design and control of electric motor and generator drives. This computational burden is particularly visible in the design of generators for wind energy harvest and motors for propulsion of electric and plug-in hybrid electric vehicles where significant field statistics are often used in the design process. Field reconstruction method (FRM) is a technique in which a minimal set of field analysis is used to establish basis functions for the magnetic flux densities in the machine. Once the basis functions are established, the performance of the machine is predicted under arbitrary speed and excitation. It has been shown that the FRM can reduce the computational time by two or three orders of magnitude while maintaining the same level of precision that is usually accomplished in a two-dimensional FE analysis. Due to its high computational efficiency, FRM has created new opportunity for health monitoring, fault-tolerant operation, and multiphysics design of electric machines. Over the past 6 years, this technique has been successfully applied to different types of machines [11]. This method still needs more researches.

The main disadvantage of PMSMs is nonuniformity in the developed torque (torque ripple) [12]. These torque pulsations vary periodically with rotor position and are reflected as speed ripple. This ripple is parasitic, and can lead to mechanical vibration, acoustic noise, and drive system problems. Therefore, minimizing the torque ripple is of great importance for PMSM design. Cogging torque is the consequence of interaction (magnetic attraction) between rotor-mounted permanent magnets field and the stator teeth, which produces reluctant variations



depending on the rotor position; it is stator current independent.

In the literature, numerous methods for reducing cogging torque, such as employing a fractional number of slots per pole, skewing of magnets or stator lamination stack (slots), displacing and shaping the magnets, optimizing magnet pole-arc-to-pole-pitch ratio, introducing auxiliary slots or teeth, etc., have been proposed.

Reducing acoustic noise in an interior permanent magnet motor during the design stage is another major issue related to IPMs [13]. Noise sources in electric machines are broadly categorized as magnetic, mechanical, electronic, and aerodynamic sources. Generally, the mechanical and magnetic sources are considered as main noise sources of electric motors because the interaction between the harmonic of magnetic forces and the resonant frequency of mechanical structure is the major cause of vibration which produces the noise. So, when designing high-speed permanent magnet (PM) machine, an effort to obtain eddy-current losses reduction in the magnets is necessary in order to improve their performances and diminish the risk of demagnetization. The iron losses in machines however, are strongly affected by the frequency of the excitation, which is proportional to the speed of the machine. In order to reduce iron losses, it is important to consider using alternative magnetic materials in the core of the machines [14]. So, optimal design of PMSMs which still is an open area of research can be translated into minimization of the loss while reducing the flux and torque ripples.

## Bibliography

### Primary Literature

- Fahimi B (2007) Qualitative approach to electromechanical energy conversion: reinventing the art of design in adjustable speed drives. In: ICEMS international conference on electrical machines and systems, Beijing, pp 432–439
- Jiang W, Moallem M, Fahimi B, Pekarek S (2006) Qualitative investigation of force density components in electromechanical energy conversion process. In: IEEE industrial electronics, IECON 2006 – 32nd annual conference, Paris, pp 1113–1118
- Arroyo ELC (2006) Modeling and simulation of permanent magnet synchronous motor drive system. Master thesis in electrical engineering, University of Puerto Rico
- Strnat KJ (1990) Modern permanent magnets for applications in electro-technology. *Proc IEEE* 78(6):923
- Fahimi B (2002) Fault tolerant motor drives. In: Jurgen RK (ed) *Handbook of automotive electronics*. CRC Press, Boca Raton
- Krause PC, Wasynczuk O, Sudhoff SD (2002) *Analysis of electric machinery and drive systems*, 2nd edn. Wiley Inter-Science, New York
- Rahman MF, Haque ME, Tang L, Zhong L (2004) Problems associated with the direct torque control of an interior permanent magnet synchronous motor drive and their remedies. *IEEE Trans Ind Electron* 51(4):799–809
- Yu H, Fahimi B (2008) A novel detection technique of non-uniform airgap in frictionless linear induction transportation systems. In: IEEE VEHICLE POWER AND PROPULSION CONFERENCE (VPPC), Harbin, 3–5 Sept 2008
- Nandi S (2006) Detection of stator faults in induction machines using residual saturation harmonics. *IEEE Trans Ind Appl* 42(5):1201–1208
- Bennani ABA, Ghodbane Cherif M, Belkhdja IS (2008) New fault tolerant DTC control for induction machine drives. In: 13th power electronics and motion control conference (EPE-PEMC), Poznan
- Fahimi B, Pekarek S (2010) Design and control of electric machines utilizing a field reconstruction method (FRM). In: 36th annual conference on IEEE industrial electronics society, IECON, Glendale
- Guemes JA, Iraolagoitia AM, Del Hoyo JI, Fernandez P (2011) Torque analysis in permanent-magnet synchronous motors: a comparative study. *IEEE Trans Energy Convers* 26(1):55–63
- Lee S, Hong J, Hwang SM, Lee WT, Lee JY, Kim YK (2009) Optimal design for noise reduction in interior permanent-magnet motor. *IEEE Trans Ind Appl* 45(6):1954–1960
- Liew GS, Ertugrul N, Soong WL, Gayler J (2006) An investigation of advanced magnetic materials for axial field brushless permanent magnet motor drives for automotive applications. In: 37th IEEE power electronics specialists conference (PESC '06), Jeju

### Books and Reviews

- Belahcen A (1999) Overview of the calculation methods for forces in magnetized iron cores of electrical machines. Presented at the Seminar on modeling and simulation of multi-technological machine systems, vol 29, pp 41–47
- Desai PC, Krishnamurthy M, Schofield N, Emadi A (2010) Novel switched reluctance machine configuration with higher number of rotor poles than stator poles: concept to implementation. *IEEE Trans Ind Electron* 57(2):469–476
- Edrington CS, Kaluvagunta DC, Joddar J, Fahimi B (2005) Investigation of electromagnetic force components in SRM under single and multiphase excitation. *IEEE Trans Ind Appl* 41(4):978–988
- Hayashi H, Nakamura K, Chiba A, Fukao T, Tungpimolrut K, Dorrell DG (2009) Efficiency improvements of switched reluctance motors with high-quality iron steel and enhanced conductor slot fill. *IEEE Trans Energy Convers* 24(4):819–825

- Holtz J (1996) Methods for speed sensorless control of AC drives. In: Rajashekara K, Kawamura A, Matsuse K (eds) Sensorless control of AC motor drives. IEEE Press, Piscataway
- Hwu KI, Liaw CM (2003) Intelligent tuning of commutation for maximum torque capability of a switched reluctance motor. *IEEE Trans Energy Convers* 18(1):113–120
- Khoobroo A, Fahimi B, Pekarek S (2008) A new field reconstruction method for permanent magnet synchronous machines. In: IECON international conference on industrial electronics, Orlando, pp 2009–2013
- Kioskeridis I, Mademlis C (2006) Optimal efficiency control of switched reluctance generators. *IEEE Trans Energy Convers* 21(4):1062–1071
- Krishnan R (2001) Switched reluctance motors drives: modeling simulation, analysis, design, and applications. CRC Press, Boca Raton
- Lin F, Yang S (2007) Instantaneous shaft radial force control with sinusoidal excitations for switched reluctance motors. *IEEE Trans Energy Convers* 22(3):629–636
- Mao SH, Tsai MC (2005) A novel switched reluctance motor with C-core stators. *IEEE Trans Magn* 41(12):3334–3336
- Mecrow C, El-Kharashi EA, Finch JW, Jack AG (2004) Preliminary performance evaluation of switched reluctance motors with segmental rotors. *IEEE Trans Energy Convers* 19(4):679–686
- Zhu W, Fahimi B, Pekarek S (2006) A field reconstruction method for optimal excitation of permanent magnet synchronous machines. *IEEE Trans Energy Convers* 21(2):305–313

## Active Multifocal Vision System, Adaptive Control of

ALOIS UNTERHOLZNER, HANS-JOACHIM WUENSCH  
Institute for Autonomous Systems Technology,  
Department of Aerospace Engineering, University of  
the Bundeswehr Munich, Neubiberg, Germany

### Article Outline

Glossary  
Definition of the Subject  
Introduction: The Human Eye as a Template  
Related Work  
Design of the Visual Sensor  
Realizing Visual Motor Routines  
Saccades and Smooth-Pursuit Movements  
Vestibulo-Ocular Reflex  
Future Directions: The Importance of Visual Perception  
Bibliography

### Glossary

**Adaptive control** The adaptation of the control mechanism to unknown or changing plant characteristics.

**Multifocal vision system** A set of cameras with different focal length. Potentially the orientation of the cameras can be actively changed. Equivalent to camera platform.

**Saccade** Rapid, abrupt sensor movement to bring an interesting object into the sensor focus.

**Selective attention** Selecting the most informative parts within the sensor's field of view for further exploration.

**Smooth-pursuit movement** Slow sensor movement to keep an object centered in the field of view.

**Vestibulo-ocular reflex** Sensor movement to compensate disturbing motions of the sensor carrier.

**Visual odometry** Determining the position and orientation of a moving camera by analyzing the camera images.

### Definition of the Subject

The perception of the environment is crucial for autonomous systems and visual perception in particular is a substantial source of information for intelligent vehicles. For a visual perception system of an intelligent vehicle high measurement accuracy combined with a large field of view is essential. Since these are contradictory requirements for a single camera, multifocal vision systems – equipped with tele- and wide-angle cameras – are frequently used. However, due to their small aperture angles high resolution telecameras have to be actively directed toward scene regions of interest. Moreover active inertial gaze stabilization is required since telecameras are highly sensitive to rotational vehicle motions. Gaze stabilization as well as adjusting the orientation of the telecamera requires control of the vision system. The use of adaptive control allows for varying camera configurations and thereby further enhances the applicability of multifocal vision systems.

### Introduction: The Human Eye as a Template

Visual perception is one of the most important senses for biological as well as for technical systems. However for technical systems it is still challenging to extract meaningful information from visual sensor input.

Due to its impressive performance in interpreting visual scenes, the human sense of vision may serve as a template for an artificial visual sense.

Similar to a camera in technical systems the eye is the sensor of the human visual system. The human eye provides high visual acuity in the central region of the retina – the fovea – close to the optical axis. Simultaneously the eye offers a wide field of view due to a large but low-resolution peripheral area surrounding the fovea. Therefore peripheral vision delivers highly compressed but less detailed data compared to the high-resolution foveal vision. Since data compression already takes place in the sensor, there is the need for a selection stage – often referred to as selective attention – that decides which parts of the environment are most informative and therefore should be discovered in more detail.

Selective attention combined with high-resolution foveal vision in turn requires directing the fovea actively toward selected regions of interest (ROI). Moreover during the measurement process the ROI shall be kept foveated despite relative motions of the ROI with respect to the sensor. Shifting the fovea toward a ROI as fast as possible is called a saccade. Smooth-pursuit movements keep the selected ROI within the fovea based on visual feedback. However visual feedback is subject to large latencies and therefore smooth-pursuit movements are inappropriate to compensate for, e.g., fast head motions. The vestibulo-ocular reflex (VOR) inertially stabilizes the image on the retina based on measurements of the vestibulo organ located in the inner ear. The combination of the vestibulo-ocular reflex and smooth-pursuit movements ensures stable tracking of selected ROIs. Binocular vision additionally requires vergence movements – simultaneous movements of both eyes in opposite directions – to keep a selected ROI foveated in both eyes.

In summary the visual properties of the human eye in combination with selective attention lead to decreased data rates with increased information density compared to a sensor with high resolution in the whole field of view.

An artificial visual sensor for intelligent vehicles mimicking the desirable visual properties of the human eye implies the need for a multifocal vision system comprising at least one tele- and one wide-angle

camera. Such a sensor should be active and provide the visual motor routines: saccade, smooth-pursuit movement, and vestibulo-ocular reflex. The need for vergence movements decreases with the increasing distance of the observed features. Therefore the capability of vergence movements is not essential for intelligent vehicles and can be omitted in favor of a considerable reduction of the mechanical complexity of the resulting multifocal vision system.

Before going into the details of the design and control of active multifocal vision systems, this entry gives a brief overview on related work. The design of visual sensors for intelligent vehicles is discussed in the subsequent section. In the following sections the realization of the visual motor routines is presented. The entry finally concludes with a discussion of possible future research directions.

## Related Work

This entry is mainly on the design and control of multifocal vision systems. Yet to provide a more general overview on the subject, literature on the human visual sense is presented first.

A profound discussion of the human visual sense including eye movements and visual attention is given in [1]. A more general treatment of human perception is presented in the textbooks [2] and [3]. Recent findings regarding human gaze control and visual attention can be found, e.g., in the reviews [4–9]. Human visual perception during driving is investigated in [10–16], for example.

Several approaches exist to the design and control of active vision systems. Most of them are more or less inspired by the human eye, mimicking to some extent its sensing properties and visual motor routines. Vision systems for robots – including humanoids – are described, e.g., in [17–27]. Simulating human binocular vision is the purpose of the vision system presented in [28]. The design and control of vision systems for autonomous ground vehicles is discussed in [29–38].

The research on computational realizations of visual attention may be divided into two – partially overlapping – groups:

The first group deals mainly with computational models of visual attention with the purpose of identifying interesting scene regions and possibly

directing a camera toward this region. These models mainly result in the computation of saliency maps based on bottom-up feature extraction and top-down weighting of these features, e.g., to find a certain object of interest. However up to now these approaches are not or rarely used for tracking applications. The publications [39–58] provide by no means a complete overview. Michalke et al. [59] present an application of this approach in the context of intelligent driver assistance systems.

In the second group the possibility of camera motions is utilized to enhance the capabilities of a robotic system. Some of the pioneering work on what is commonly referred to as “active perception” or “active vision” is presented in [60–64]. Dickmanns [65] presents an early approach to vision-based vehicle guidance. A lot of research is concerned with active vision or gaze control respectively to improve self-localization or obstacle avoidance of (humanoid) robots: [66–76]. There are also some works on simultaneous localization and mapping (SLAM): [77–79]. Gaze control for autonomous vehicles is presented, e.g., in [80–82] and [29]. Most of these works are based on simulations or were conducted in comparably simple structured and almost static environments.

### Design of the Visual Sensor

Active multifocal vision systems – or equivalently camera platforms – are able to provide simultaneously a wide field of view, high resolution, and stereo vision.

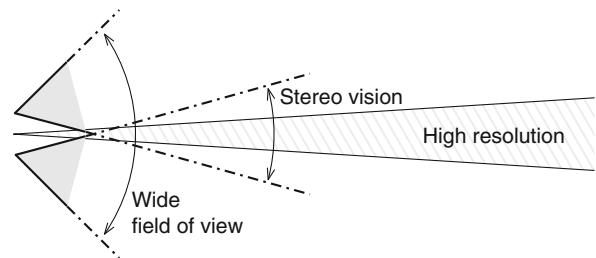
A wide field of view is desirable to scan simultaneously as much of the vehicle’s environment as possible. A high resolution provides detailed images of points of interest in the vehicle’s surrounding. Moreover telecameras are required for early detection of distant objects, since the maximum range of object detection limits the maximum speed of the vehicle. Stereo vision is used to gather 3D information of the environment.

Active vision systems that are pivotable in pitch and yaw axis enable the exploitation of the high measurement accuracy offered by telecameras within the entire field of view of the vision system. Moreover the visual field of the intelligent vehicle is enlarged this way. However telecameras are more sensitive to

disturbances than wide-angle cameras and therefore active gaze stabilization is required to obtain stable images and to keep objects of interest within camera scope.

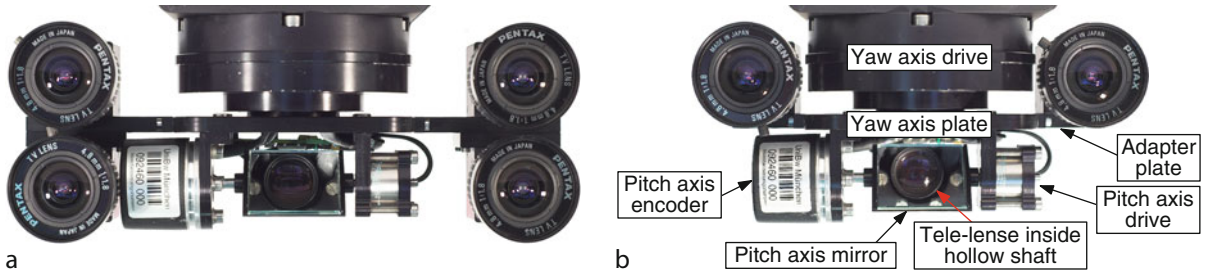
For a sound discussion of these requirements and the resulting design of multifocal vision systems refer to [29] or [30]. According to [30], platforms pivotable in pitch and yaw axis and equipped with several cameras, that either provide a wide field of view or a high resolution, offer a good compromise to meet the needs of visual perception in autonomous ground vehicles. Figure 1 depicts one possible multifocal camera arrangement following the recommendations given in [29].

Although multifocal vision systems offer a good compromise covering more or less the needs of each visual perception module of an intelligent vehicle, the optimal camera configuration for a single perception module can differ from the currently utilized camera arrangement. To detect, e.g., lane markings black/white cameras might be best; to detect drivable areas in off-road terrain color cameras might be better. Divergent cameras offer a wide field of view. Parallel cameras, on the contrary, increase the range where stereo vision is possible. Depending on their range of interest different perception modules require different focal lengths of camera lenses. Therefore software designers prefer to use the camera configuration that suits best their application’s demands. For this reason the configuration of the multifocal vision system of an intelligent vehicle may be frequently changed during the development process.



**Active Multifocal Vision System, Adaptive Control of.**  
**Figure 1**

A multifocal vision system offering a wide field of view combined with a high-resolution central area. Stereo vision is possible within the intersection of the field of view of the two wide-angle cameras



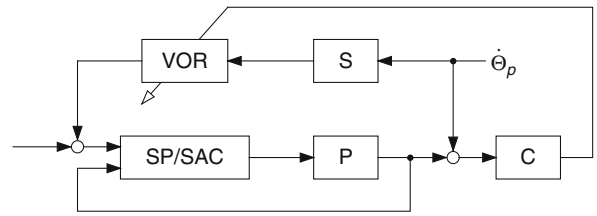
**Active Multifocal Vision System, Adaptive Control of. Figure 2**  
Two possible configurations of the proposed multifocal vision system

With every change of the cameras the extrinsic camera parameters are lost and recalibration is required. Moreover exchanging cameras may alter camera platform dynamics due to different camera masses, which in turn alters platform inertia leading to reduced motion control performance of the multifocal vision system.

To overcome these problems the presented multifocal vision system not only follows the design recommendations given in [29], but additionally offers the possibility to exchange cameras while retaining camera calibration for later reuse of this configuration as suggested in [83]. Two possible system configurations are shown in Fig. 2. For easy exchange of the cameras they are mounted on adapter plates (Fig. 2b). These plates are attached to the yaw axis plate. For high repeat accuracy the plates are positioned with adjust pins relative to the yaw axis plate. Thus camera calibration is preserved as long as the cameras remain fixed to the adapter plates. Pitch axis drive and position sensor are mounted below the yaw axis plate. The pitch axis camera is placed in the hollow shaft of the yaw axis drive. A mirror is used to reflect the horizontal view toward the camera in the hollow shaft. Inertia of the pitch axis is kept low this way since only the mirror has to be rotated. Thus the weight of the pitch axis drive can also be kept to a minimum, which in turn reduces the inertia of the yaw axis. Placing the pitch axis camera in the center of rotation of the yaw axis further reduces yaw axis inertia.

### Realizing Visual Motor Routines

An active multifocal vision system implies the need for visual motor routines like saccades, smooth-pursuit movement, and the vestibulo-ocular reflex.



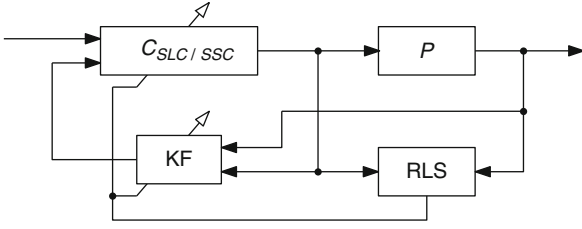
**Active Multifocal Vision System, Adaptive Control of. Figure 3**  
Control of the multifocal vision system comprising three distinct controllers for saccades, smooth-pursuit movements, and the vestibulo-ocular reflex

Saccades are fast movements to a given reference position. Smooth-pursuit movements are used to track objects of interest or equivalently to keep selected regions of interest foveated. Saccades and smooth-pursuit movements require reference values either from a single visual perception module, e.g., a vehicle tracker or – based on the current situational context – from an attention module. Both saccades (SAC) and smooth-pursuit movements (SP) are based on feedback control that directs the plant  $P$  – which is the multifocal vision system – to a given reference position (Fig. 3).

The vestibulo-ocular reflex (VOR) provides active gaze stabilization based on inertial measurements of disturbing platform motions. A MEMS (micro-electro-mechanical system) sensor  $S$  is placed on the yaw axis plate of the multifocal vision system to measure the angular velocity  $\dot{\theta}_p$  of the pitch axis induced by vehicle rotations. Visual feedback from the camera  $C$  is used to calibrate this open-loop gaze stabilization.

Due to the different control objectives, the realization of saccades, smooth-pursuit movements, and the





**Active Multifocal Vision System, Adaptive Control of.**  
**Figure 4**

Block diagram control of saccade and smooth pursuit

vestibulo-ocular reflex requires three distinct controllers (Fig. 3). The controllers for saccades and smooth-pursuit movements are implemented for both the yaw and the pitch axis of the multifocal vision system. The vestibulo-ocular reflex is only realized for the platform pitch axis, since disturbances like bumps, braking, or accelerating mainly induce vehicle pitch rotations.

### Saccades and Smooth-Pursuit Movements

Saccades can be considered to be a time-optimal control problem. Therefore a near time-optimal sliding-mode control law ( $C_{SLC}$ ) is proposed for saccades (Fig. 4). The smooth-pursuit movements are governed by a state-space controller ( $C_{SSC}$ ) that is designed to provide good tracking behavior. Both controllers share a Kalman filter (KF) to derive full state information and a recursive least squares (RLS) parameter estimator for online identification of the parameters of the dynamic model of plant  $P$ . The RLS estimator is necessary since altering the camera configuration of the multifocal vision system may lead to a change in the platform dynamics – which is obvious from Fig. 2a and b. The parameter estimates of the RLS estimator are used to adapt the Kalman filter KF as well as the control laws  $C_{SLC}$  and  $C_{SSC}$ . Based on the control error a switch determines which control law is applied.

### Dynamic Model

The dynamic plant models of platform yaw and pitch axis differ only slightly. Since the pitch axis model emerges from the yaw axis model, the yaw axis model is derived first. Plant dynamics is modeled as a nonlinear second-order differential equation as in [83] and [84]:

$$J \cdot \ddot{\varphi} = k_m \cdot I - \tau_\varphi - \tau_\varphi \quad (1)$$

$I$  is the torque generating motor current,  $k_m$  the motor torque constant, and  $J$  the inertia of the platform yaw axis. The angular velocity-dependent disturbance torque  $\tau_\varphi$  is described by

$$\tau_\varphi = \tau_c \cdot \text{sgn}(\dot{\varphi}) + d \cdot \dot{\varphi} \quad (2)$$

with  $\tau_c$  representing Coulomb and  $d$  representing viscous friction coefficient, with  $\text{sgn}(\dot{\varphi})$  being the signum function. The angular position-dependent disturbance torque  $\tau_\varphi$  results from twisted cables and from the unbalanced mass  $m$  with respect to the platform yaw axis:

$$\tau_\varphi = r \cdot m \cdot g \cdot \sin(\theta_p) \cdot \cos(\varphi) + c \cdot \varphi \quad (3)$$

The torsion-spring model  $c \cdot \varphi$  describes the torque resulting from twisted cables. The term  $r \cdot m \cdot g \cdot \sin(\theta_p) \cdot \cos(\varphi)$  represents the disturbance torque resulting from the unbalanced mass, with  $r$  being the distance from the center of gravity of the unbalanced mass  $m$  to the center of rotation of the platform yaw axis,  $g$  being the gravitational acceleration, and  $\theta_p$  being the adjustable pitch angle of the platform yaw axis. The unbalanced mass results from the pitch axis encoder outweighing the pitch axis drive and any mass differences of the used camera/lens combinations. Defining  $r \cdot m \cdot g \cdot \sin(\theta_p) =: \tau_u$ , the disturbance torque  $\tau_\varphi$  can be written as

$$\tau_\varphi = \tau_u \cdot \cos(\varphi) + c \cdot \varphi \quad (4)$$

With (2) and (4) the plant dynamics (1) can be summarized to:

$$J \cdot \ddot{\varphi} = k_m \cdot I - d \cdot \dot{\varphi} - c \cdot \varphi - \tau_c \cdot \text{sgn}(\dot{\varphi}) - \tau_u \cdot \cos(\varphi) \quad (5)$$

The dynamic plant model of the platform pitch axis can be derived from the yaw axis model by neglecting the angular position-dependent disturbance torque (4) in (5).

### Parameter Estimation

Since the model parameters  $J$ ,  $d$ ,  $c$ ,  $\tau_c$  and  $\tau_u$  are unknown, a parameter estimator is used to get the estimates  $\hat{J}$ ,  $\hat{d}$ ,  $\hat{c}$ ,  $\hat{\tau}_c$ , and  $\hat{\tau}_u$  respectively. Discretizing (5) using the forward differences as suggested in [85]



$$\dot{\varphi} = (\varphi_{k+1} - \varphi_k)/T \quad (6)$$

a discrete time plant model

$$\begin{aligned} \varphi_k = & -a_1 \cdot \varphi_{k-1} - a_2 \cdot \varphi_{k-2} + b_2 \cdot I_{k-2} \\ & - \delta_1 \cdot \operatorname{sgn}\left(\frac{\varphi_{k-1} - \varphi_{k-2}}{T}\right) - \delta_2 \cdot \cos(\varphi_{k-2}) \end{aligned} \quad (7)$$

is derived which is linear in the parameters:

$$\begin{aligned} a_1 = \frac{Td}{J} - 2, \quad a_2 = 1 - \frac{Td}{J} + \frac{cT^2}{J}, \quad b_2 = \frac{k_m T^2}{J} \\ \delta_1 = \frac{\tau_c T^2}{J}, \quad \delta_2 = \frac{\tau_u T^2}{J} \end{aligned} \quad (8)$$

Then (7) can be written as linear regression which is suitable for the application of a RLS parameter estimator. Parameter projection is used to constrain the parameter estimates to stay within reasonable limits as outlined in [86] and [83].

### State Estimation

Both state-space control and sliding-mode control require full state information. Therefore, analog to [83] and [84], a state estimator is applied, using the angular position  $\varphi$  as measurement input. Since the plant model (5) is nonlinear, the parameter estimates  $\hat{\tau}_c$  and  $\hat{\tau}_u$  are used to compensate for the nonlinear disturbance terms in (5) with the disturbance feedforward:

$$I_{ff} = k_m^{-1} \cdot (\hat{\tau}_c \cdot \operatorname{sgn}(\dot{\varphi}) + \hat{\tau}_u \cdot \cos(\varphi)) \quad (9)$$

In [85], a similar approach is suggested for adaptive friction compensation. The plant input current  $I$  can be written as:

$$I = I_{ff} + I_{fb} + I_{ref} \quad (10)$$

with  $I_{ff}$  resulting from feedforward control,  $I_{fb}$  resulting from feedback control, and  $I_{ref}$  resulting from the reference input  $r$  of the axis controller. By inserting feedforward control (9) in the plant model (5) the equation

$$\begin{aligned} J \cdot \ddot{\varphi} = & k_m \cdot (I_{fb} + I_{ref}) - d \cdot \dot{\varphi} - c \cdot \varphi \\ & - (\tau_c - \hat{\tau}_c) \cdot \operatorname{sgn}(\dot{\varphi}) - (\tau_u - \hat{\tau}_u) \cdot \cos(\varphi) \end{aligned} \quad (11)$$

is derived.

Summarizing the remaining disturbance torques resulting from parameter estimation errors  $\tau_u - \hat{\tau}_u$  and  $\tau_c - \hat{\tau}_c$ , as well as from unmodeled disturbances in one disturbance torque  $\tau_d$  (11) can be written as

$$J \cdot \ddot{\varphi} = k_m \cdot (I_{fb} + I_{ref}) - d \cdot \dot{\varphi} - c \cdot \varphi - \tau_d \quad (12)$$

$\ddot{\tau}_d = 0$  is used as a disturbance model, as this is found in [87] to be a good compromise between model complexity and estimation accuracy in the context of friction compensation.

Using the continuous time state-space formulation

$$\begin{aligned} \dot{x} &= Fx + G\bar{u}, \quad y = Cx, \quad \text{with} \\ x &= (\varphi, \dot{\varphi}, \tau_d, \dot{\tau}_d)^T, \quad \bar{u} = I_{fb} + I_{ref} \end{aligned} \quad (13)$$

from (12) and disturbance model  $\ddot{\tau}_d = 0$  the state-space description

$$\begin{aligned} F &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ -\frac{c}{J} & -\frac{d}{J} & -\frac{1}{J} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad G = \begin{pmatrix} 0 \\ \frac{k_m}{J} \\ 0 \\ 0 \end{pmatrix}, \\ C &= (1, 0, 0, 0) \end{aligned} \quad (14)$$

can be derived. As a discrete time Kalman filter is used for state estimation, a discrete time description

$$\begin{aligned} x_{k+1} &= Ax_k + B\bar{u}_k, \quad y_k = Cx_k, \quad \text{with} \\ x_k &= (\varphi_k, \dot{\varphi}_k, \tau_{dk}, \dot{\tau}_{dk})^T \end{aligned} \quad (15)$$

of (14) is needed. Using (6) for discretization yields

$$\begin{aligned} A &= \begin{pmatrix} 1 & T & 0 & 0 \\ -\frac{cT}{J} & \frac{1-Td}{J} & -\frac{T}{J} & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ \frac{k_m T}{J} \\ 0 \\ 0 \end{pmatrix} \\ C &= (1, 0, 0, 0) \end{aligned} \quad (16)$$

Replacing parameters  $J$ ,  $c$ , and  $d$  at every time step  $k$  with their estimates  $\hat{J}_k$ ,  $\hat{c}_k$ , and  $\hat{d}_k$  results in  $\hat{A}_k$  and  $\hat{B}_k$  as state and input matrices respectively.

### State-Space Control

State-space control is used for smooth-pursuit movements that demand a good tracking behavior. Therefore the reference  $r$  is designed to track the desired axis angle  $\varphi_{ref}$  up to a ramp signal without steady-state error. Thus  $r$  is a linear combination of  $\varphi_{ref}$  and its

time derivative  $\dot{\varphi}_{ref}$  with gains  $w_1$  and  $w_2$ :

$$r = w_1 \cdot \varphi_{ref} + w_2 \cdot \dot{\varphi}_{ref} \quad (17)$$

Following the controller design in [83] and [84], the steady-state error of the discrete time tracking error  $\varphi_{ref}(z) - \varphi(z)$  is calculated by applying the final value theorem. Provided  $\hat{c}$  and  $\hat{d}$  approach their true values, the steady-state tracking error becomes zero if  $w_1$  and  $w_2$  are chosen as follows:

$$w_1 = k_1 + \frac{\hat{c}}{k_m} \text{ and } w_2 = k_2 + \frac{\hat{d}}{k_m} \quad (18)$$

Using the estimated state vector  $\hat{x}_k$  and the disturbance feedforward  $I_{ff}$  together with the discrete time reference input  $r_k$  yields the control law:

$$u_k = r_k - k_{1k} \hat{\varphi}_k - k_{2k} \dot{\hat{\varphi}}_k + \frac{\hat{\tau}_{d_k}}{k_m} + \frac{\hat{\tau}_{c_k}}{k_m} \text{sgn}(\dot{\hat{\varphi}}_k) + \frac{\hat{\tau}_{u_k}}{k_m} \cos(\hat{\varphi}_k) \quad (19)$$

Applying the control law (19) to the plant (5), results in a second-order closed-loop system. As described in [83], the controller gains  $k_1$  and  $k_2$  are determined by the comparison of coefficients with the desired second-order closed-loop characteristic polynomial. The derived gains are

$$k_1 = (\hat{J}\omega_0^2 - \hat{c})/k_m \text{ and } k_2 = (2\zeta\omega_0\hat{J} - \hat{d})/k_m \quad (20)$$

with closed-loop natural frequency  $\omega_0$  and damping  $\zeta$ . Damping is chosen  $\zeta > 1$  to avoid overshoot and  $\omega_0$  is calculated as follows [83, 88]:

$$\omega_0 = \frac{4.6}{\eta\zeta\sqrt{\hat{J}}} \quad (21)$$

The constant  $\eta$  is used to adjust the influence of the estimated inertia  $\hat{J}$  on the desired closed-loop dynamics  $\omega_0$ .

### Sliding-Mode Control

Saccadic movements are a typical application for time-optimal control. Time-optimal control problems can be solved using Pontryagin's minimum principle [89], but the solution is sensitive to unknown plant

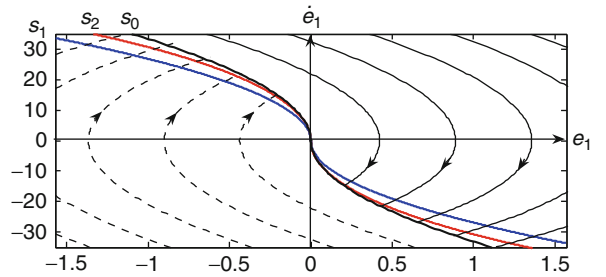
dynamics or disturbances. Moreover it is difficult to derive a solution analytically for nonlinear systems.

Sliding-mode control on the contrary is an inherently robust method to cope with uncertain plant dynamics. If the controlled system has reached the sliding-mode, it is insensitive to disturbances and system motion is determined by the sliding-surface only [90, 91].

There are several approaches that try to utilize the robustness properties of sliding-mode control, while simultaneously attaining near time-optimal control performance, e.g., [92–94].

These control law design methods are offline methods and therefore the derived controllers are not suitable for online adaptation. The design of the presented saccade controller is based on the minimum time sliding-mode control law for second-order systems derived in [94]. However the sliding-line proposed in [94] is subject to stability constraints that prevent the control from being time-optimal. Analog to [84], the proposed sliding-line is designed to relax these constraints and thus achieve near time optimality.

The system motion of a system under sliding-mode control consists of two modes: the *reaching-mode* and the *sliding-mode* [91]. During the *reaching-mode* the phase trajectory of the system (Fig. 5) starts from anywhere in the phase plane and moves toward the sliding-line  $s = 0$ . After reaching the sliding-line  $s = 0$  the phase trajectory will stay on it, provided the *reaching condition* is satisfied. Then the system is in the so-called sliding-mode and the system motion is only governed by the differential equation  $s = 0$ . With the system under control



Active Multifocal Vision System, Adaptive Control of.

Figure 5

Phase trajectories and sliding-lines [84]

$$u = -K \operatorname{sgn}(s), K > 0 \quad (22)$$

and using the Lyapunov function candidate

$$V(s) = \frac{1}{2}s^2 \quad (23)$$

a global reaching condition [91, 94] is given by:

$$\dot{V}(s) = s\dot{s} < 0, s \neq 0 \quad (24)$$

That means  $s > 0$  indicates  $\dot{s} < 0$  and vice versa. Thus the phase trajectories will always move toward and stay on  $s = 0$  after reaching it.

In order to apply sliding-mode control, plant dynamics (5) is rewritten as a set of first-order differential equations. Defining the control error  $\varphi - \varphi_{ref} := e_1$  and noting that  $\varphi_{ref}$  is constant yields

$$\begin{aligned} \dot{e}_1 &= e_2 \\ \dot{e}_2 &= (f(\dot{\varphi}) + g(\varphi) + h(\varphi, \dot{\varphi}) + u)/J \end{aligned} \quad (25)$$

with  $\dot{e}_1 = \dot{\varphi}$  and  $\dot{e}_2 = \ddot{\varphi}$ . The disturbance torque is separated into three categories:  $f(\dot{\varphi})$  describes the friction torque, with the property  $f > 0$  if  $\dot{\varphi} < 0$  and vice versa. Secondly  $g(\varphi)$  is a spring torque and is assumed to be monotonically decreasing, i.e.,  $\partial g / \partial \varphi < 0 \forall \varphi$ . Finally,  $h(\varphi, \dot{\varphi})$  sums up the remaining disturbance torques that cannot be described by  $f$  or  $g$ .  $u$  is the plant input torque.

For a second-order system the nonlinear switching surface

$$s = \alpha e_1 + |e_2|e_2 = 0 \quad (26)$$

is proposed in [94]. Solving the differential equation (26) for constant  $\alpha$ , the time span  $t$  needed to approach  $e_1 = 0$  is calculated as follows:

$$t = 2\sqrt{|e_1(t=0)|/\alpha} \quad (27)$$

Provided that (24) is satisfied, it is obvious from (27) that the resulting transient response is faster for higher  $\alpha$ . Thus  $\alpha$  has to be maximized to reach near time-optimal control. For a double-integrator plant the sliding-line (26) becomes identical to the time-optimal switching-line for  $\alpha \rightarrow 2K$  [94], with  $K$  being the maximum control torque. In [94] it is argued that if the plant is not a pure double integrator but subject to disturbances as, e.g., system (25), the maximum values of the disturbances have to be considered in the

calculation of  $\alpha$ . Otherwise there are no resources left to compensate for the disturbances. Following this argumentation, it can be shown that the *reaching condition* (24) is satisfied for system (25) with control (22) and switching surface (26) for

$$\alpha < \frac{2}{J}(K - f_{max} - g_{max} - h_{max}) \quad (28)$$

But supposing that, e.g.,  $f$  describes viscous friction  $f = d\dot{\varphi}$ ,  $f_{max}$  depends linearly on the largest possible angular velocity  $\dot{\varphi}_{max}$ . Thus  $\alpha$  is too conservative for  $\dot{\varphi} \ll \dot{\varphi}_{max}$ .

The proposed approach utilizes the properties of the disturbance torques  $f$  and  $g$  to increase  $\alpha$ , leading to a faster transient response of the control loop:

$$\alpha < \frac{2}{J}(K + g \cdot \operatorname{sgn}(e_1) - h_{max}) \quad (29)$$

The term  $g \cdot \operatorname{sgn}(e_1)$  is chosen such that  $\alpha$  increases if  $g$  reduces acceleration or supports deceleration. On the contrary  $\alpha$  decreases if  $g$  supports acceleration or reduces deceleration. The friction torque  $f$  always counteracts the actuator torque and can thus be neglected in the calculation of  $\alpha$ . Therefore  $\alpha$  is independent from  $f_{max}$  and  $g_{max}$ , which is a desirable property. Figure 5 depicts some switching surfaces of system (25) with  $\varphi_{ref} = 0$ . The time-optimal switching-line is  $s_0$ .  $s_1$  and  $s_2$  are the switching-lines from (26) with  $\alpha$  according to (28) or (29) respectively.  $s_2$  lies in between  $s_0$  and  $s_1$ , indicating that using (29) to calculate  $\alpha$  yields a faster transient response than using (28).

Comparing system (25) with the plant dynamics (5), one derives

$$\begin{aligned} f &= -d \cdot \dot{\varphi} - \tau_c \cdot \operatorname{sgn}(\dot{\varphi}) \\ g &= -c \cdot \varphi \\ h &= -\tau_u \cdot \cos(\varphi) \end{aligned} \quad (30)$$

for the disturbance torques  $f$ ,  $g$ , and  $h$ . The maximum control torque  $K$  in (22) corresponds to the maximum actuator torque  $k_m \cdot I_{max}$ .

Since the plant parameters are not known in advance, the parameter estimates derived from the RLS parameter estimator are used (cf. section “[Parameter Estimation](#)”) for the calculation of the sliding-line. Moreover the state estimates are obtained from the Kalman filter introduced in section “[State Estimation](#).” Thus the sliding-line is

$$s = \alpha(\hat{\varphi}_k - \varphi_{ref}) + |\dot{\hat{\varphi}}_k| \dot{\hat{\varphi}}_k \quad (31)$$

with

$$\alpha < \frac{2}{f} \left( K - \hat{c} \cdot \hat{\varphi}_k \cdot \text{sgn}(\hat{\varphi}_k - \varphi_{ref}) - |\hat{\tau}_u| \right) \quad (32)$$

### Switching Strategy

To switch between state-space control and sliding-mode control the simple hysteresis

$$\begin{aligned} \text{IF } |\varphi_{ref} - \varphi| > \varepsilon_u \text{ THEN saccade} &= \text{true} \\ \text{IF } |\varphi_{ref} - \varphi| > \varepsilon_l \text{ THEN saccade} &= \text{false} \end{aligned}$$

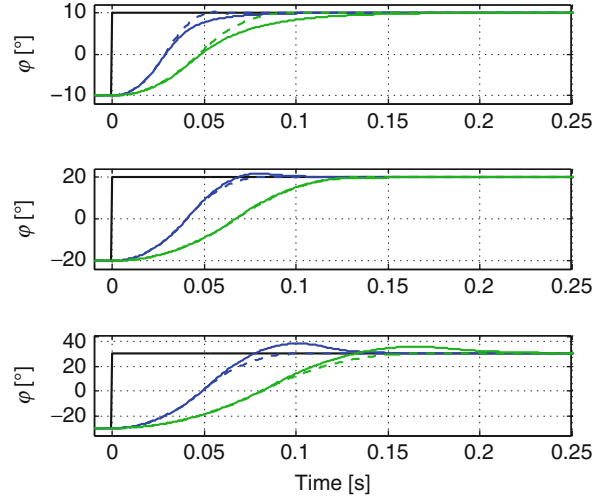
is used, with  $\varepsilon_u > \varepsilon_l > 0$ . If the control error  $\varphi_{ref} - \varphi$  exceeds the upper threshold  $\varepsilon_u$ , a saccade is carried out and sliding-mode control is active. If the control error goes below the lower threshold  $\varepsilon_l$ , state-space control becomes active until the control error exceeds  $\varepsilon_u$  again.

### Experimental Results

Some experimental results described in [84] demonstrate the validity of the presented saccade and smooth-pursuit control laws. The two configurations of the multifocal vision system shown in Fig. 2 were used for the experiments. These configurations equipped with three (Fig. 2b) and five cameras (Fig. 2a) are called 3c- and 5c-configuration respectively. Saccades with three different step sizes were conducted. For every step size the results of the sliding-mode control are compared with the state-space control, this way demonstrating the superior performance of the proposed near time-optimal sliding-mode control over state-space control with respect to saccades. The step sizes were  $\pm 10^\circ$ ,  $\pm 20^\circ$ , and  $\pm 30^\circ$  (Fig. 6). Every plot of Fig. 6 depicts four saccades: two for each platform configuration, resulting from sliding-mode control and state-space control respectively.

For fast saccades maximum acceleration followed by maximum deceleration is required. Sliding-mode control design is explicitly based on the maximum achievable acceleration and deceleration. That is why sliding-mode control suits saccadic movements better than state-space control.

From Fig. 6 it is obvious that saccades under state-space control are either too slow or result in overshoot.



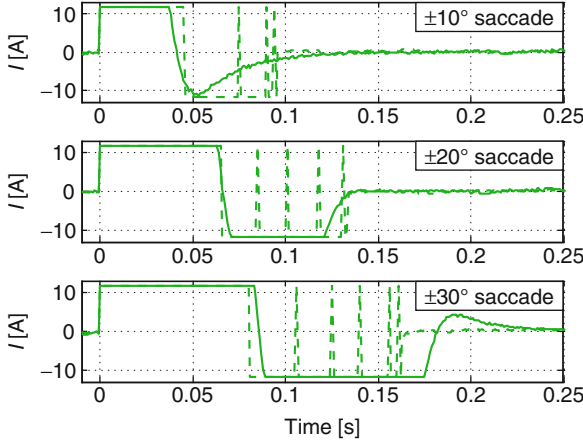
**Active Multifocal Vision System, Adaptive Control of.**  
**Figure 6**

Saccades [84]: Black lines are reference values. Blue lines correspond to the 3c-configuration, green lines to the 5c-configuration. The dashed lines result from the sliding-mode control, the solid lines from state-space control

The results for sliding-mode and state-space control are similar only for a small operating range as can be seen from the  $\pm 20^\circ$ -saccade.

Figure 7 depicts the controller output corresponding to the saccades of the 5c-configuration shown in Fig. 6. For the  $\pm 20^\circ$ -saccade the controller output of state-space control and sliding-mode control overlap to a large extent. In particular the switching from acceleration to deceleration and vice versa takes place at nearly identical points in time. Therefore the resulting saccades are also similar. During the  $\pm 30^\circ$ -saccade the state-space control output changes from acceleration to deceleration later than the sliding-mode control output. Also the deceleration phase is finished later, resulting in considerable overshoot. For the  $\pm 10^\circ$ -saccade the state-space controller starts decelerating earlier than the sliding-mode controller and thus the corresponding saccade is finished later.

The switching activity (chattering) of the sliding-mode controller is low, indicating that the control performance is close to time-optimal. For practical applications, however, it is preferable to avoid chattering completely, since it may cause increased hardware attrition and induces audible noise.



**Active Multifocal Vision System, Adaptive Control of.**  
**Figure 7**

Controller output for saccades of the 5c-configuration [84]: Solid lines correspond to state-space control, dashed lines to sliding-mode control

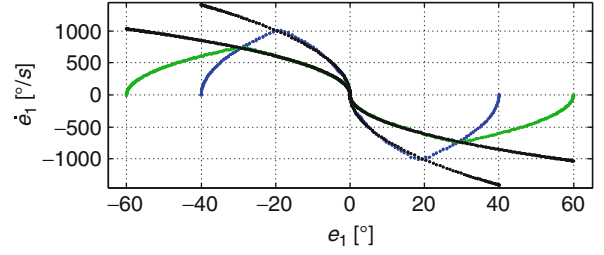
Replacing the signum function in (22) with the saturation function

$$u = K \text{sat}(s) = \begin{cases} -\frac{K}{l}s & \text{for } s \leq |l| \\ -K \text{sgn}(s) & \text{for } s > |l| \end{cases} \quad (33)$$

is a possible alternative. The saturation function introduces a boundary layer around the sliding-line  $s$ . Outside the boundary layer the saturation function equals the signum function. Inside the boundary layer the sliding-mode control is replaced by a high-gain proportional control, thereby avoiding the discontinuity of the signum function for  $s = 0$ .

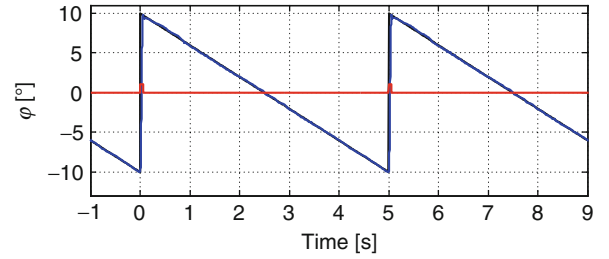
The sliding-lines for the 3c- and 5c-configuration are shown in Fig. 8. Note that the slope of the sliding-line of the 3c-configuration is higher than the slope of the 5c-configuration. This results mainly from the influence of the estimated inertia  $\hat{J}$  on the calculation of  $\alpha$  in (32). Obviously the inertia of the 3c-configuration is smaller than the inertia of the 5c-configuration.

Moreover the resulting phase trajectories of the  $\pm 20^\circ$ -saccade for the 3c-configuration and the  $\pm 30^\circ$ -saccade for the 5c-configuration are depicted in Fig. 8. After reaching the sliding-line, the trajectories stay in a small vicinity of the sliding-line, due to the nonideal components of a real-world control loop.



**Active Multifocal Vision System, Adaptive Control of.**  
**Figure 8**

Sliding surfaces and phase trajectories [84]: The blue line corresponds to the phase trajectory of the 3c-configuration, the green line to the trajectory of the 5c-configuration. The black lines are the corresponding sliding surfaces



**Active Multifocal Vision System, Adaptive Control of.**  
**Figure 9**

Saccades and smooth-pursuit movements [84]: The black line is the reference value. The blue line corresponds to the 3c-configuration. The red line states which control law is active: state-space or sliding-mode control

A sawtooth signal is used as position reference to test the switching between sliding-mode control and state-space control (Fig. 9). With the sawtooth signal saccades and smooth-pursuit movements can be simulated since it offers fast (saccade like) changes as well as continuous change over time with a constant slope. The output of the switching hysteresis proposed in section “Switching Strategy” is depicted in red. Both smooth-pursuit movements and saccades show satisfactory control behavior. Furthermore the switching between both controllers is smooth and without noticeable influence on the control performance.

## Vestibulo-Ocular Reflex

The use of high resolution telecameras enhances the perceptual capabilities of autonomous vehicles. It enables both the early detection of distant objects and provides detailed images of points of interest in the vehicles surrounding. But telecameras are more sensitive to disturbances – especially rotational vehicle motion – than wide-angle cameras.

To obtain stable images and keep objects of interest within camera scope active gaze stabilization is necessary. In [29, 33], inertial measurements of the rotational vehicle motion are used for active image stabilization in autonomous vehicles.

This is a pure feedforward control, since image-based information is not used for the stabilization. Residual image motion may be induced by errors of the inertial angular-rate sensor, such as sensor offset, wrong or changing sensor sensitivity, errors in the mounting position of the sensor, or imperfections in the control loop of the stabilizing actuator.

Therefore in more recent works visual information is incorporated in the stabilization process. Either information from inertial sensors and vision is fused as, e.g., in [24], or the inertial sensors are calibrated based on visual information as in [27, 36, 37]. The latter approach is similar to the vestibulo-ocular reflex (VOR) that many vertebrates use for gaze stabilization. The VOR not only stabilizes the image on the retina, but also continuously adapts to erroneous measurements of the vestibular organ located in the inner ear [37]. This adaptation process is based on visual feedback.

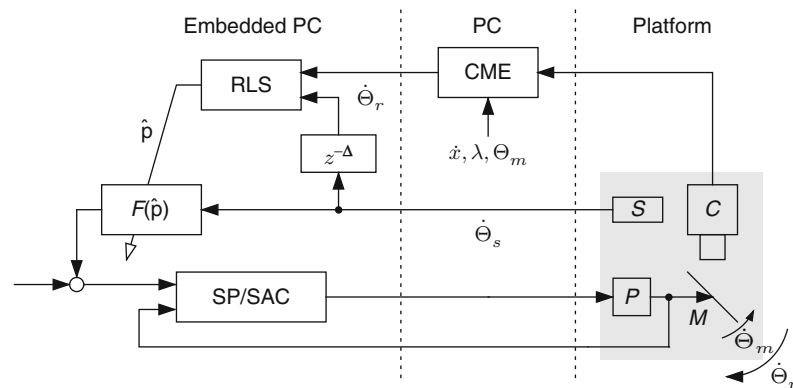
In [27, 36, 37] sensor and actuator errors are compensated by means of an adaptive FIR (finite impulse response) filter as suggested in [95]. The filter adaptation is driven by optical flow. The camera platform in [27] executes only rotational motions in an indoor environment. Thus optical flow results solely from rotational motion. In [36, 37] the camera platform is mounted behind the windshield of a car. Optical flow computation is based on distant features. This way the translational platform motion has little impact on the resulting optical flow. But calculating the optical flow by utilizing distant features only is not always possible, consider, e.g., a car driving on a forest track or in an inner-city environment.

The presented active inertial gaze stabilization is inspired by the vestibulo-ocular reflex too. In contrast to the aforementioned works, however, the sensor and actuator errors are modeled explicitly. From those error models a filter transfer function is derived that is suitable to modify the sensor signal such as to compensate for the overall stabilization error. The adaptation of the unknown filter parameters is driven by the residual camera motion.

## Calibrating the Inertial Gaze Stabilization

Analog to [96], the realization of the vestibulo-ocular reflex includes the open-loop gaze stabilization based on an MEMS angular-rate sensor and the calibration of this sensor driven by visual feedback (Fig. 10).

The angular-rate sensor  $S$  measures the pitch velocity  $\dot{\Theta}_p$  of the camera platform. After passing the filter



Block diagram of the inertial gaze stabilization and its calibration



$F(\hat{\mathbf{p}})$  the sensor signal  $\dot{\Theta}_s$  is the reference value for the control (SP/SAC) of the gaze stabilizing plant  $P$ . A mirror together with an actuator is used for gaze stabilization. In fact the reference value for the control of  $P$  is half the size of the filtered sensor signal, since for a mirror the incident angle equals the emergent angle.

Image correspondences computed from already stabilized images from camera  $C$  together with the vehicle velocity  $\dot{x}$  and the vehicle steering angle  $\lambda$  serve as measurements for the camera motion estimation (CME) based on an unscented Kalman filter (UKF) [97]. The CME is used to estimate the residual pitch velocity  $\dot{\Theta}_r$ , which is the difference between the real platform velocity  $\dot{\Theta}_{pl}$  and two times the mirror velocity  $\dot{\Theta}_m$ .  $\dot{\Theta}_r$  again is used as an error signal for a recursive least squares (RLS) parameter estimator that estimates the parameters  $\hat{\mathbf{p}}$  of the filter  $F(\hat{\mathbf{p}})$  in a way that the residual image motion is minimized. To consider the latencies  $\Delta$  resulting from image processing in the parameter estimation process, the sensor signal  $\dot{\Theta}_s$  is artificially delayed by  $\Delta$  before entering the RLS parameter estimator.

### Compensating the Open-Loop Errors

To derive an adaptive filter that compensates for the open-loop errors of the inertial gaze stabilization, the sensor and actuator characteristics are modeled explicitly as in [96].

The MEMS angular-rate sensor signal

$$\dot{\Theta}_s = k_s \dot{\Theta}_{pl} + \delta_s \quad (34)$$

is considered to be biased resulting from an offset  $\delta_s$ . Moreover the sensor is thought to have a sensitivity  $k_s$  differing from the nominal sensitivity. Please note that the sensitivity  $k_s$  may also be influenced by the mounting position of the sensor, e.g., when the sensor is not exactly aligned with the axis of rotation of the mirror.

The stabilizing mirror and its control are modeled as a first-order transfer function in the Laplace domain

$$\dot{\Theta}_m = \frac{0.5k_a}{\tau_a s + 1} \dot{\Theta}_s \quad (35)$$

with the time constant  $\tau_a$  describing the bandwidth of the control loop and the gain  $k_a$ . The gain  $k_a$  describes the steady-state deviation of the mirror velocity  $\dot{\Theta}_m$  from half of the measured sensor value  $\dot{\Theta}_s$ .

The goal of the adaptive filter design is to reconstruct the real platform pitch velocity  $\dot{\Theta}_p$  from the sensor signal  $\dot{\Theta}_s$  and to compensate the influence of the actuator control loop on the gaze stabilization. The filter equation is chosen as follows:

$$\dot{\Theta}_F = F(\hat{\mathbf{p}}) = \frac{\hat{\tau}_a s + 1}{\hat{k}_a \hat{k}_s} (\dot{\Theta}_s - \hat{\delta}_s) \quad (36)$$

Using the output of filter (36) as input to the mirror control loop (35) and inserting (34) in (36) results in

$$\dot{\Theta}_m = \frac{0.5k_a}{\tau_a s + 1} \frac{\hat{\tau}_a s + 1}{\hat{k}_a \hat{k}_s} (k_s \dot{\Theta}_p + \delta_s - \hat{\delta}_s) \quad (37)$$

If the estimated parameters  $\hat{\mathbf{p}} = [\hat{\tau}_a, \hat{k}_a, \hat{k}_s, \hat{\delta}_s]^T$  approach their true values  $\mathbf{p} = (\tau_a, k_a, k_s, \delta_s)^T$  (37) reduces to

$$\dot{\Theta}_m = 0.5 \dot{\Theta}_p \quad (38)$$

which is the desired optimal gaze stabilization.

To get estimates  $\hat{\mathbf{p}}$  of the unknown parameters  $\mathbf{p}$  with a discrete time RLS parameter estimator (Fig. 10), the filter equation (36) has to be discretized. Since the sampling interval  $T$  is small compared to the vehicle's pitch dynamics, backward differences  $\dot{x}(k) = \frac{x(k) - x(k-1)}{T}$  can be used for discretization:

$$\dot{\Theta}_F(k) = \frac{\hat{\tau}_a + T}{\hat{k}_a \hat{k}_s T} \dot{\Theta}_s(k) - \frac{\hat{\tau}_a}{\hat{k}_a \hat{k}_s T} \dot{\Theta}_s(k-1) - \frac{\hat{\delta}_s}{\hat{k}_a \hat{k}_s} \quad (39)$$

The resulting discretized filter equation (39) is linear in its parameters and can be written as a linear regression

$$\begin{aligned} \dot{\Theta}_F(k) &= [\dot{\Theta}_s(k), \dot{\Theta}_s(k-1), 1]^T \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{d} \end{bmatrix} \\ &=: \phi^T(k-1) \hat{\theta} \end{aligned} \quad (40)$$

which is suitable for the RLS parameter estimation [86].

Usually the RLS parameter update is driven by the difference of a measured signal  $y$  and a prediction  $\hat{y}$  of  $y$  based on the estimated parameters  $\hat{\theta}$ . The difference  $y - \hat{y}$  is a measure for the RLS parameter estimation error. The residual camera pitch velocity  $\dot{\Theta}_r$  can also be considered to be a measure for the parameter

estimation error. Unless  $\dot{\Theta}_r = 0$  the estimated parameters  $\hat{\mathbf{p}}$  of filter  $F(\hat{\mathbf{p}})$  are not equal to their true values  $\mathbf{p}$ . Therefore equation

$$\hat{\theta}(k+1) = \hat{\theta}(k) + K(k) \dot{\Theta}_r \quad (41)$$

can be used for the parameter update, with  $K(k)$  being the parameter update gain of the RLS parameter estimator.

### Camera Motion Estimation

The estimation of the unknown residual camera motion (CME) is based on image data. This is commonly referred to as *visual odometry*. The estimation of non-holonomic vehicle motion in real time was presented, e.g., in [98, 99]. An overview on the relevant literature is also given therein. *Visual odometry* relies on the calculation of image correspondences between consecutive image frames. A method to calculate these correspondences in real time is given in [100].

Most works on *visual odometry* aim at estimating the camera motion from visual data only. However to enhance the robustness of the CME the vehicle velocity and the vehicle steering angle are incorporated in the estimation process as suggested in [96]. These measurements are available on most ground-moving vehicles.

One fundamental problem in *visual odometry* is the determination of the depth of image keypoints. To circumvent this, the drivable area in front of the vehicle is assumed to be an ideal ground plane. Considering only image correspondences corresponding to that ground plane implicitly determines the depth of the image keypoints due to the known camera position relative to the ground plane.

As a tribute to computation speed, the calculated image correspondences include some *false positives* [100]. Thus a preselection of valid correspondences is necessary; otherwise the estimation performance may be corrupted.

The gaze stabilization is independent of the camera motion estimation. Only the calibration of the angular-rate sensor is based on the estimated camera motion. But the calibration can be switched off, whenever the estimated camera motion seems to be unreliable. This may be the case, e.g., if too little image correspondences are available or if the ground plane assumption is invalid.

The calibration of the gaze stabilization is driven by the estimated residual camera motion. However this estimation is not the main purpose of this entry. The interested reader may refer to [96] for a more detailed description of the applied camera motion estimation or to the literature on *visual odometry* for alternative approaches.

### Experimental Results

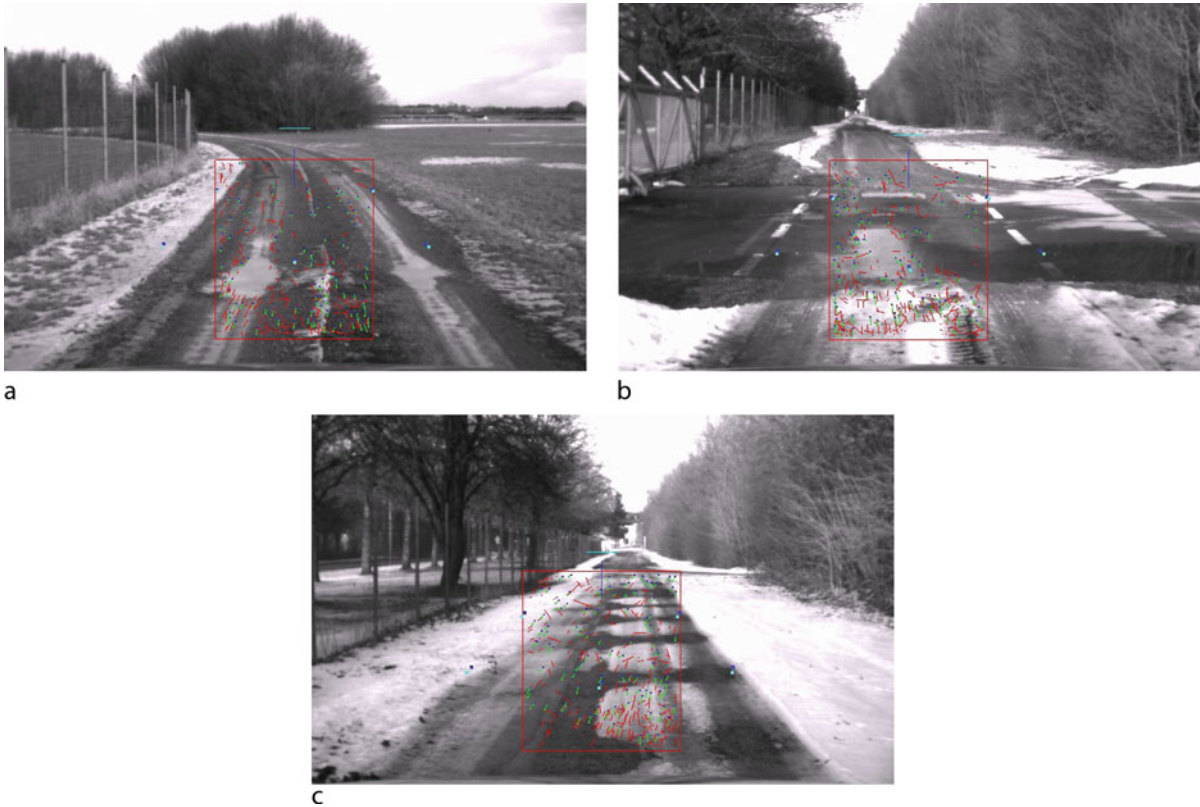
The proposed gaze stabilization was evaluated during several test runs on a bumpy dirt road as described in [96]. Figure 11 gives some impressions of the test track.

The red rectangles in Fig. 11a–c mark the region of interest of the feature correspondence algorithm. Feature correspondences are only computed within this region. According to the ground plane assumption the ROI corresponds to the drivable area in front of the vehicle. The green lines in Fig. 11a–c mark correspondences that were accepted as measurements for the camera motion estimation. The red lines depict correspondences that were considered to be false positives.

The proposed vision-based online-calibration of the MEMS angular-rate sensor was tested during three consecutive runs along the test track with activated sensor calibration and gaze stabilization. Figure 12 depicts the estimated gain  $\hat{k}_f = \frac{1}{k_a k_s}$  of the filter  $F(\hat{\mathbf{p}})$ . For evaluation purposes the filter gain was artificially reduced from 1 to 0.5. During parameter estimation the gain rose from  $\hat{k}_f = 0.5$  to  $\hat{k}_f \approx 0.88$ , which is beneath the nominal gain 1.

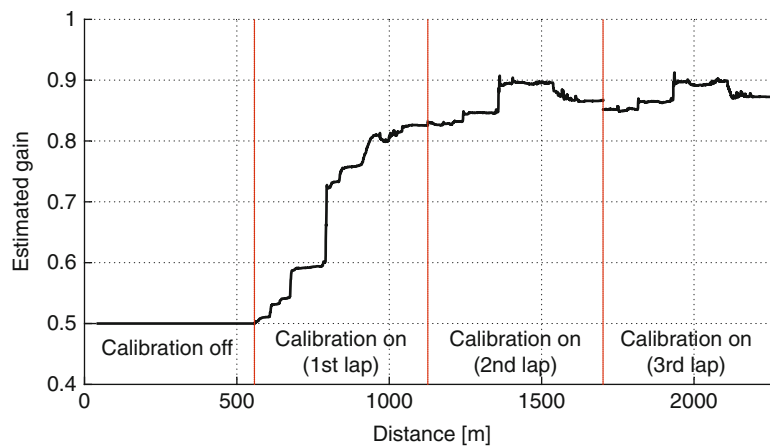
The estimated time constant  $\hat{\tau}_a$  remained close to its initial value zero. This indicates that the bandwidth of the actuator control loop is sufficiently large and no phase-delay occurs during stabilization. The estimated residual pitch velocity  $\dot{\Theta}_r$  was biased and had to be high-pass filtered to avoid parameter drift, therefore the offset  $\delta_s$  was estimated to be zero too.

To further evaluate the proposed gaze stabilization, the estimated residual camera motion  $\dot{\Theta}_r$  was compared for the following three runs on the test track: One with deactivated gaze stabilization, another one with activated stabilization but uncalibrated sensor ( $\hat{k}_f = 1$ ), and finally the one with activated gaze stabilization and calibrated sensor ( $\hat{k}_f \approx 0.88$ ). During this last run the calibration was active, but the parameters had already converged.



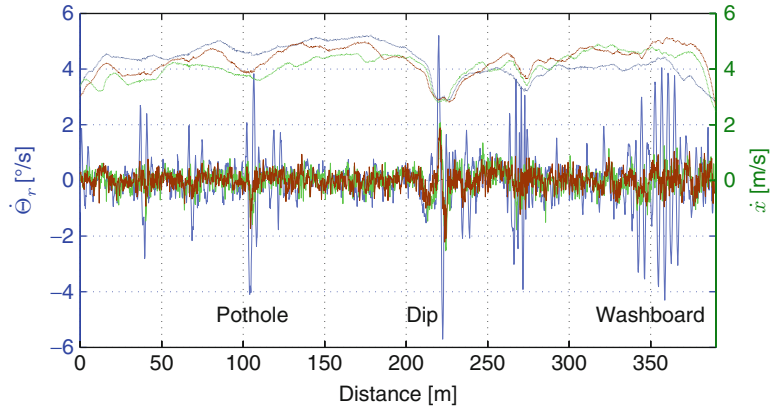
Active Multifocal Vision System, Adaptive Control of. Figure 11

Selected images from the test track [96]



Active Multifocal Vision System, Adaptive Control of. Figure 12

Estimated gain of the filter  $F(\mathbf{p})$  [96]



**Active Multifocal Vision System, Adaptive Control of. Figure 13**

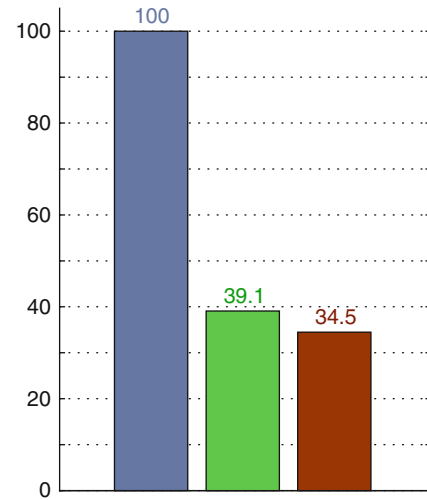
Gaze stabilization results [96]: The blue lines correspond to the deactivated stabilization, the green lines to the activated stabilization, and the dark red lines to the calibrated stabilization. Solid lines correspond to the residual camera motion  $\dot{\Theta}_r$ , dotted lines to the vehicle velocity  $\dot{x}$

The stabilization results are plotted in Fig. 13. It gives an overview of the resulting estimated camera pitch velocities  $\dot{\Theta}_r$  during the test runs. The labels “pothole,” “dip,” and “washboard” mark peaks in the pitch velocities  $\dot{\Theta}_r$  that originate from passing the track sections depicted in Fig. 11a–c.

The gaze stabilization based on the uncalibrated MEMS sensor already showed good results. Therefore for better evaluation the rms-value of  $\dot{\Theta}_r$  was calculated as a measure for the stabilization error (Fig. 14). The resulting rms-value was normalized with respect to the rms-value of  $\dot{\Theta}_r$  in case of the unstabilized camera. The normalized rms-error for the unstabilized camera is then 100%. It turns out that the rms-error corresponding to the calibrated gaze stabilization is smaller than the rms-error corresponding to the uncalibrated gaze stabilization.

The rms-error for the calibrated stabilization is approximately 12% smaller than the rms-error resulting from the uncalibrated stabilization. This can be considered to be the average reduction of the stabilization error, resulting from an increasing or a decreasing filter gain  $\hat{k}_f$ . The reduction of the rms-error corresponds to the magnitude of change of the filter gain.

The test runs were carried out at similar but not equal velocities (Fig. 13). However, the velocity during the test run with calibrated stabilization (dark red) was almost always higher than the velocity



**Active Multifocal Vision System, Adaptive Control of. Figure 14**

The normalized rms-error of the gaze stabilization (based on  $\dot{\Theta}_r$ ) [96]. The blue bar corresponds to the deactivated stabilization, the green bar to the activated stabilization, and the dark red bar to the calibrated stabilization

corresponding to the test run with active but uncalibrated stabilization (green). Although higher vehicle velocities are assumed to induce higher camera pitch velocities, the calibrated stabilization still beats the uncalibrated stabilization.

## Future Directions: The Importance of Visual Perception

Perception of the environment for intelligent vehicles should cover large parts of their surrounding with sufficient accuracy. One possible realization for the visual sensory system is a multifocal vision system simultaneously offering a wide field of view and high resolution. Obviously such a system has to be active, since the high visual acuity is limited to the small aperture angles of the tele-lenses. Saccades, smooth-pursuit movements, and the vestibulo-ocular reflex are therefore essential capabilities to be able to fully exploit the possibilities offered by an active multifocal vision system.

However multifocal vision systems are only beneficial if there is a competition for sensing resources, be it either to observe characteristic parts of an object in more detail or to observe regions of interest that are not in the current sensor view. Then, similar to human selective attention, a selection stage is necessary that decides which gaze direction is most appropriate to the task at hand.

The outcome of most current computational models of visual attention are saliency maps of interesting scene regions. These maps can be used to direct the telecamera of a multifocal vision system toward this salient region to obtain high-resolution images thereof. If top-down cues are incorporated in the map calculation they can be used, e.g., for object detection of previously known objects. However, if at all, up to now saliency maps are only rarely used for object tracking.

Multifocal object tracking requires object recognition based on the identification and tracking of a set of characteristic object features, e.g., the lights, the tires, or the license plate of a car. After identifying these features in the image frame of a wide-angle camera they could be successively observed with a high resolution tele-lense. This kind of object tracking is subject to current research and therefore it is developed for a single camera first.

Apart from the computation of saliency maps, there exist some approaches to gaze control that are used to track multiple objects, landmarks, or road networks. These approaches either try to predict the most informative viewing direction(s) based on information theoretic concepts or to plan a sequence of gaze directions

ensuring an optimal viewing behavior considering the needs of several distributed visual perception modules with respect to the visibility of interesting scene regions. However these existing gaze control approaches were not employed in dynamic environments like inner-city traffic scenes.

Providing visual perception for intelligent vehicles, capable of reliable and robust multiple object tracking together with the detection and tracking of intersections, roads, traffic signs, and so on, is a major challenge in future research. Multifocal object tracking and computational models of visual attention may contribute to this by no means trivial task. Such advanced perceptual capabilities give rise to a competition for sensing resources. Then the intelligent selective gaze control – based on the current context – choosing the appropriate gaze directions will be inevitable.

## Bibliography

1. Palmer SE (1999) Vision science. MIT Press, Cambridge, MA
2. Kolb B, Whishaw IQ (2009) Fundamentals of human neuropsychology, 6th edn. Worth, New York
3. Kandel ER, Schwartz JH, Jessell TH (1995) Essentials of neural science behavior, 1st edn. Prentice Hall International, London
4. Henderson JM (2003) Human gaze control during real-world scene perception. *Trends Cogn Sci* 7(11):498–504, November
5. Treue S (2003) Visual attention: the where, what, how and why of saliency. *Curr Opin Neurobiol* 13:428–432
6. Chun MM, Marois R (2002) The dark side of visual attention. *Curr Opin Neurobiol* 12:184–189
7. Krauzlis RJ (2005) The control of voluntary eye movements: new perspectives. *Neuroscientist* 11(2):124–137
8. Knudsen EI (2007) Fundamental components of attention. *Annu Rev Neurosci* 30:57–78
9. Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. *Trends Cogn Sci* 13(9):403–409
10. Land MF, Lee DN (1994) Where we look when we steer. *Nature* 369:742–744
11. Chapman PR, Underwood G (1998) Visual search of driving situations: danger and experience. *Perception* 27:951–964
12. Olson PL, Battle DS, Aoki T (1989) Driver eye fixations under different operating conditions, University of Michigan, Ann Arbor, Transportation Research Institute, Technical Report
13. Wann JP, Wilkie RM (2004) How do we control high speed steering. In: *Optic flow and beyond*. Kluwer, Norwell, pp 401–419, ch 18
14. Falkmer T, Gregersen NP (2005) A comparison of eye movement behavior of inexperienced and experienced drivers in real traffic environments. *Optom Vis Sci* 82(8):732–739



15. Metz B (2009) Worauf achtet der Fahrer? steuerung der aufmerksamkeit beim fahren mit visuellen nebenaufgaben. Ph.D. dissertation, Julius-Maximilians-Universität Würzburg
16. Doshi A, Trivedi MM (2009) On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Trans Intell Transp Syst* 10(3):453–462
17. Kühnlenz K (2007) "Aspects of multi-focal vision," ser. Mess-, Steuerungs- und Regelungstechnik.VDI Verlag 8(1129)
18. Kühnlenz K, Bachmayr M, Buss M (2006) A multi-focal high-performance vision system. In: Proceedings of the 2006 IEEE international conference on robotics and automation, Orlando, pp 150–155
19. Ude A, Gaskett C, Cheng G (2006) Foveated vision systems with two cameras per eye. In: Proceedings of the 2006 IEEE international conference on robotics and automation (ICRA 2006), Orlando
20. Roca X, Vitrià J, Vanrell M, Villanueva JJ (1999) Gaze control in a binocular robot systems. In: Proceedings of the 1999 IEEE international conference on emerging technologies and factory automation (ETFA 1999), Spain
21. Peng J, Srikaew A, Wilkes M, Kawamura K, Peters A (2000) An active vision system for mobile robots. In: Proceedings of the 2000 IEEE international conference on systems, man and cybernetics (ICSMC 2000), Nashville
22. Yamato J (1999) A layered control system for stereo vision head with vergence. In: Proceedings of the 1999 IEEE international conference on systems, man and cybernetics (ICSMC 1999), Tokyo
23. Panerai F, Metta G, Sandini G (2000) Adaptive image stabilization: a need for vision-based active robotic agents. In: Proceedings of the international conference on the simulation of adaptive behavior (SAB 2000), Paris
24. Lobo J, Ferreira JF, Dias J (2008) Robotic implementation of biological Bayesian models towards visuo-inertial image stabilization and gaze control. In: Proceedings of the 2008 IEEE international conference on robotics and biomimetics, Bangkok
25. Suppa M, Kielhöfer S, Langwald J, Hacker F, Strobl KH, Hirzinger G (2007) The 3D-modeller: a multi-purpose vision platform. In: Proceedings of the 2007 IEEE international conference on robotics and automation (ICRA 2007), Roma
26. Vilgratner T, Ulbrich H (2011) Design and control of a compact high-dynamic camera-orientation system. *IEEE-ASME T Mech* 16(2):221–231
27. Lenz A, Balakrishnan T, Pipe AG, Melhuish C (2008) An adaptive gaze stabilization controller inspired by the vestibulo-ocular reflex. *Bioinspir Biomim* 3(3):035001
28. Zhang X, Tay ALP (2009) A physical system for binocular vision through saccade generation and vergence control. *Cybernet Syst* 40:549–568
29. Dickmanns ED (2007) Dynamic vision for perception and control of motion. Springer, London
30. Dickmanns ED (2003) An advanced vision system for ground vehicles. In: Proceedings of the 1st international workshop on in-vehicle cognitive computer vision systems, Graz
31. Schiehlen J, Dickmanns ED (1994) A camera platform for intelligent vehicles. In: Proceedings of the 1994 IEEE intelligent vehicles symposium (IV 1994), Paris
32. Schiehlen J, Dickmanns ED (1994) Design and control of a camera platform for machine vision. In: Proceedings of the 1994 IEEE international conference on intelligence in robotics and autonomous systems (IROS 1994), Neubiberg
33. Schiehlen J (1995) Kameraplattformen für aktiv sehende fahrzeuge ser. Mess-, Steuer- und Regelungstechnik. VDI Verlag 8(514)
34. Günthner W, Wagner P, Ulbrich H (2006) An inertially stabilised vehicle camera system – hardware, algorithms, test drives. In: Proceedings of the 2006 IEEE international conference on industrial electronics (IECON 2006), Paris
35. Günthner W, Glasauer S, Wagner P, Ulbrich H (2006) Enhancing inertial measurement units by multi-sensor fusion with visual data for active camera stabilization. In: Proceedings of joint conference of the 37th international symposium on robotics ISR 2006 and the German conference on robotics, Munich
36. Guenther W, Glasauer S, Wagner P, Ulbrich H (2007) Biomimetic control for adaptive camera stabilization in driver-assistance systems. *J Mech Sci Technol* 21(6):930–934
37. Günthner W (2008) Enhancing cognitive assistance systems with inertial measurement units. Springer, Berlin
38. Ducrocq Y, Bahrami S, Duvieubourg L, Cabestaing F (2008) An effective active vision system for gaze control. In: Advances in visual computing, ser. Lecture notes in computer science, vol 5359. Springer, Berlin, pp 267–277
39. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol* 4:219–227
40. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell* 20(11):1254–1259
41. Driscoll JA, Peters RA II, Cave KR (1998) A visual attention network for a humanoid robot. In: Proceedings of the 1998 IEEE/RSJ international conference on intelligent robotic systems (IROS 1998), Victoria
42. Backer G, Mertsching B, Bollmann M (2001) Data- and model-driven gaze control for an active-vision system. *IEEE Trans Pattern Anal Mach Intell* 23(12):1415–1429
43. Breazeal C, Scassellati B (1999) A context-dependent attention system for a social robot. In: Proceedings of the sixteenth international joint conference on artificial intelligence (IJCAI 1999), Stockholm
44. Vijayakumar S, Conradt J, Shibata T, Schaal S (2001) Overt visual attention for a humanoid robot. In: Proceedings of the 2001 IEEE international conference on intelligence in robotics and autonomous systems (IROS 2001), Hawaii
45. Sprague N, Ballard D (2003) Eye movements for reward maximization. In: Advances in neural information processing systems 15. MIT Press, Cambridge, MA
46. Choi SB, Jung BS, Ban SW, Lee M (2004) Trainable attention model based vergence control for active stereo vision system. In: Proceedings of the 2004 IEEE international conference on



- intelligent sensors, sensor networks and information processing (ISSNIP 2004), Melbourne
47. Navalpakkam V, Itti L (2005) Modeling the influence of task on attention. *Vis Res* 45:205–231
  48. Orabona F, Metta G, Sandini G (2005) Object-based visual attention: a model for a behaving robot. In: *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR 2005)*, San Diego
  49. Ude A, Wyart V, Lin LH, Cheng G (2005) Distributed visual attention on a humanoid robot. In: *Proceedings of the 2005 IEEE-RAS international conference on humanoid robots*, Tsukuba
  50. Frintrop S (2006) VOCUS: a visual attention system for object detection and goal-directed search. *Lecture notes in computer sciences ed., ser. (Lecture notes in artificial intelligence)*, vol 3899. Springer, Heidelberg
  51. Begum M, Mann GKI, Gosine RG (2006) A biologically inspired Bayesian model of visual attention for humanoid robots. In: *Proceedings of the 2006 IEEE-RAS international conference on humanoid robots (HUMANOIDS 2006)*, Genova
  52. Yu Y, Mann GKI, Gosine RG (2008) An object-based visual attention model for robots. In: *Proceedings of the 2008 IEEE international conference on robotics and automation (ICRA 2008)*, Pasadena
  53. Vogel J, de Freitas N (2007) Target-directed attention: sequential decision-making for gaze planning. In: *Proceedings of the 2007 IEEE international conference on robotics and automation (ICRA 2007)*, Pasadena
  54. Borji A, Ahmadabadi MN, Araabi BN (2009) Learning sequential visual attention control through dynamic state space discretization. In: *Proceedings of the 2009 IEEE international conference on robotics and automation (ICRA 2009)*, Kobe
  55. Xu T, Kühnlenz K, Buss M (2008) Information-based gaze control adaptation to scene context for mobile robots. In: *Proceedings of the 2008 IEEE international conference on pattern recognition (ICPR 2008)*, Tampa
  56. Xu T, Wu H, Zhang T, Kühnlenz K, Buss M (2009) Environment adapted active multi-focal vision system for object detection. In: *Proceedings of the 2009 IEEE international conference on robotics and automation (ICRA 2009)*, Kobe
  57. Itti L, Baldi P (2009) Bayesian surprise attracts human attention. *Vis Res* 49(10):1295–1306, June
  58. Xu T, Zhang T, Kühnlenz K, Buss M (2010) Attentional object detection with an active multi-focal vision system. *Int J Hum Robot* 7(2):223–243
  59. Michalke T, Kastner R, Adamy J, Bone S, Waibel F, Gayko MKA, Gepperth A, Fritsch J, Goerick C (2008) An attention-based system approach for scene analysis in driver assistance. *Automatisierungstechnik* 56(11):575–584
  60. Aloimonos JY, Weiss I, Bandyopadhyay A (1988) Active vision. *Int J Comput Vision* 7:333–356
  61. Bajcsy R (1988) Active perception. *Proc IEEE* 76(8):966–1005, Invited Paper
  62. Aloimonos JY (1990) Purposive and qualitative active vision. In: *Proceedings of the 1990 IEEE international conference on pattern recognition*, Atlantic
  63. Ballard DH (1991) Animate vision. *Artif Intell* 48:57–86
  64. Aloimonos Y (ed) (1993) *Active perception*. Lawrence Erlbaum, Hillsdale
  65. Dickmanns ED (1988) An integrated approach to feature based dynamic vision. In: *Proceedings of the 1988 IEEE computer society conference on computer vision and pattern recognition (CVPR 1988)*, Ann Arbor
  66. Davison AJ (1998) *Mobile robot navigation using active vision*. PhD Thesis, University of Oxford
  67. Davison AJ, Murray DW (1998) Mobile robot localisation using active vision. In: *Proceedings of the 5th European conference on computer vision (ECCV 1998)*, Freiburg
  68. Seara JF, Strobl KH, Schmidt G (2002) Information management for gaze control in vision guided biped walking. In: *Proceedings of the 2002 IEEE international conference on intelligent robots and systems (IROS 2002)*, Lausanne
  69. Seara JF, Strobl H, Schmidt G (2003) Path-dependent gaze control for obstacle avoidance in vision guided humanoid walking. In: *Proceedings of the 2003 IEEE international conference on robotics and automation (ICRA 2003)*, Taipei
  70. Seara JF, Strobl KH, Martín E, Schmidt G (2003) Task-oriented and situation-dependent gaze control for vision guided humanoid walking. In: *Proceedings of the 2003 IEEE international conference on humanoid robots (HUMANOIDS 2003)*, Munich
  71. Seara JF (2004) *Intelligent gaze control for vision-guided humanoid walking*, ser. Mess-, Steuer- und Regelungstechnik.VDI Verlag 8(1044)
  72. Seara JF, Schmidt G (2005) Gaze control strategy for vision-guided humanoid walking. *Automatisierungstechnik* 53(2):49–58
  73. Lidoris G, Kühnlenz K, Wollherr D, Buss M (2007) Combined trajectory planning and gaze direction control for robotic exploration. In: *Proceedings of the 2007 IEEE international conference on robotics and automation (ICRA 2007)*, Rome
  74. Kühnlenz K, Lidoris G, Wollherr D, Buss M (2007) On foveated gaze control and combined gaze and locomotion planning. In: *de Pina ACF (ed) Humanoid robots: new developments*. I-Tech Education and Publishing, Vienna
  75. Kühnlenz K, Buss M (2008) Multi-focal vision and gaze control improve navigation performance. *J Hum* 1(1):33–44
  76. Xu T, Kühnlenz K, Buss M (2008) A view direction planning strategy for a multi-camera vision system. In: *Proceedings of the 2008 IEEE international conference on information and automation (ICIA 2008)*, Zhangjiajie
  77. Davison AJ, Kita N (2001) 3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition (CVPR 2001)*, Lihue
  78. Davison AJ, Murray DW (2002) Simultaneous localization and map-building using active vision. *IEEE Trans Pattern Anal Mach Intell* 24:865–880
  79. Lidoris G, Kühnlenz K, Wollherr D, Buss M (2006) Information-based gaze direction planning algorithm for slam.

- In: Proceedings of the 2006 IEEE-RAS international conference on humanoid robots (HUMANOIDS 2006), Genova
80. Pellkofer M, Dickmanns ED (2000) EMS-vision: gaze control in autonomous vehicles. In: Proceedings of the 2000 IEEE intelligent vehicles symposium (IV 2000), Dearborn
  81. Pellkofer M (2003) Verhaltensentscheidung für autonome fahrzeuge mit blickrichtungssteuerung. Dissertation, Universität der Bundeswehr München
  82. Dickmanns ED (2004) Three-stage visual perception for vertebrate-type dynamic machine vision. In: Proceedings of the conference on engineering of intelligent systems (EIS 2004), Madeira
  83. Unterholzner A, Wuensche HJ (2009) Adaptive state space control of a camera platform for an autonomous ground vehicle. In: Proceedings of the 2009 IEEE intelligent vehicles symposium (IV 2009), Xi'an
  84. Unterholzner A, Wuensche HJ (2010) Hybrid adaptive control of a multi-focal vision system. In: Proceedings of the 2010 IEEE intelligent vehicles symposium (IV 2010), San Diego
  85. Canudas C, Åström KJ, Braun K (1987) Adaptive friction compensation in DC-motor drives. IEEE J Robot Automat RA-3(6):681–685
  86. Goodwin GC, Sin KS (1984) Adaptive filtering prediction and control, Information and system sciences. Prentice-Hall, Englewood Cliffs
  87. Ramasubramanian A, Ray LR (2007) Comparison of EKBF-based and classical friction compensation. J Dyn Syst Meas Contr 129:236–242
  88. Franklin GF, Powell JD, Emami-Naeini A (2002) Feedback control of dynamic systems, 4th edn. Prentice Hall, Upper Saddle River
  89. Athans M, Falb PL (1966) Optimal control: an introduction to the theory and its applications. McGraw-Hill, New York
  90. Utkin VI (1977) Variable structure systems with sliding modes. IEEE Trans Autom Control AC-22(2):212–222
  91. Hung JY, Gao W, Hung JC (1993) Variable structure control: a survey. IEEE Trans Ind Electron 40(1):2–22, February
  92. Chatchanayuenyong T, Parnichkun M (2008) Time optimal hybrid sliding mode-PI control for an autonomous underwater robot. Int J Adv Rob Syst 5(1):91–98
  93. You KH, Lee EB (2000) Robust, near time-optimal control of nonlinear second order systems with model uncertainty. In: Proceedings of the 2000 IEEE international conference on control applications, Anchorage, Alaska
  94. Iliev B, Kalaykov I (2004) Minimum-time sliding mode control for second-order systems. In: Proceedings of the 2004 American control conference, Boston
  95. Dean P, Porrill J, Stone JV (2002) Decorrelation control by the cerebellum achieves oculomotor plant compensation in simulated vestibulo-ocular reflex. Proc R Soc Lond Biol Sci 269(1503):1895–1904, August
  96. Unterholzner A, Rohland M, Schweitzer M, Wuensche HJ (2010) Vision based online-calibration of inertial gaze stabilization. In: Proceedings of the 2010 IEEE intelligent vehicles symposium (IV 2010), San Diego
  97. Julier SJ, Uhlmann JK (2004) Unscented filtering and nonlinear estimation. Proc IEEE 92(3):410–422, Invited Paper
  98. Schweitzer M, Wuensche HJ (2010) Real-time visual odometry for ground moving robots using GPUs. In: Proceedings of the international conference on computer vision theory and applications (VISAPP 2010), Angers
  99. Scaramuzza D, Fraundorfer F, Pollefeys M, Siegwart R (2009) Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In: Proceedings of the IEEE international conference on computer vision (ICCV 2009), Kyoto
  100. Schweitzer M, Wuensche HJ (2009) Efficient keypoint matching for robot vision using GPUs. In: Proceedings of the fifth IEEE workshop on embedded computer vision at ICCV 2009, Kyoto

## Active Pedestrian Protection System, Scenario-Driven Search Method for

ALBERTO BROGGI<sup>1</sup>, PIETRO CERRI<sup>1</sup>, STEFANO GHIDONI<sup>2</sup>,  
PAOLO GRISLERI<sup>1</sup>, HO GI JUNG<sup>3</sup>

<sup>1</sup>VisLab – Dipartimento di Ingegneria  
dell'Informazione, Università degli Studi di Parma,  
Parma, Italy

<sup>2</sup>Intelligent Autonomous System Laboratory  
(IAS-Lab), Department of Information Engineering,  
University of Padua, Padova, Italy

<sup>3</sup>The School of Mechanical Engineering, Hanyang  
University, Seondong-gu, Seoul, South Korea

### Article Outline

Glossary  
Definition of the Subject and Its Importance  
Introduction  
Why Are ADAS Related to Sustainability?  
Innovation of This Method  
State of the Art  
Setup  
Focus of Attention  
Vision Fusion  
Automatic Braking  
Results  
Future Directions  
Bibliography

### Glossary

**ADAS** *Advanced Driver Assistance Systems* are complex systems to help the driver during specific maneuvers, or during normal driving.

**Laser scanner** Laser scanner is an optical sensor based on *LIDAR* (Light Detection And Ranging) technology that measures the distance of targets measuring the time delay between transmission of a pulse and detection of the reflected signal.

**NIR** Near InfraRed (0.75–1.4  $\mu\text{m}$ ) is the wavelength closest to visible that is often used in image processing for night vision, together with specific illuminators.

### Definition of the Subject and Its Importance

This paper presents an application of a pedestrian detection system aimed at localizing potentially dangerous situations in specific urban scenarios. The approach used in this work differs from the ones implemented in traditional pedestrian detection systems, which are designed to localize all pedestrians in the area in front of the vehicle. Conversely this approach searches for pedestrians in critical areas only.

The great advantages of such an approach are that pedestrian recognition is performed on limited image areas – therefore boosting its time-wise performance – and no assessment on the danger level is finally required before providing the result to either the driver or an onboard computer for automatic maneuvers. A further advantage is the drastic reduction of false alarms, making this system robust enough to control nonreversible safety systems.

### Introduction

This paper presents an innovative approach to the detection of pedestrians, explicitly designed to work in a particularly challenging urban scenario, which represents a particularly critical environment for traditional pedestrian detection approaches. The aim of the system is to trigger reversible and nonreversible driving assistance system.

This work was developed by VisLab at Parma University and funded by MANDO Corporation, Korea.

### Why Are ADAS Related to Sustainability?

In his book *May Contain Nuts* [1] John O’Farrell depicts a mother terrified by modern world’s dangers. In order to show how dangerous the road where

vehicles are running can be, she builds a puppet, and throws it onto the road, letting it come out between two parked cars.

The description of that particular “experiment” intended to highlight two points, among others: first of all, the (motivated) worries people have about the road environment, especially when dealing with children, which cannot correctly evaluate risks; the second observation is that road safety relies, in some cases, on a very strong hypothesis, that is, everyone is following road traffic laws. Of course, this should normally be the case, but an important point is to understand whether there are margins, in order to recover for other road users’ illegal behaviors. This applies mainly to drivers, since they are the strong road users. When a car hits a pedestrian, even if this is due to a pedestrian fault, the driver could, in principle, take actions that may avoid or, at least, mitigate the collision consequences. This is, then, the difference. It is difficult, if not impossible, for a pedestrian to take actions that can recover from a driver’s fault; on the contrary, it is often possible for a driver to cope with pedestrians’ faults.

Road dangers are caused by several factors. At the very beginning, vehicles were slow, and few of them were running on roads – traffic also was an unknown problem! Then, the number of vehicles dramatically increased, but not enough care was dedicated to the protection of pedestrians and bicycles mobility, either because of economic factors, or because the problem itself is really difficult to face.

Whatever the reason, traffic and related pedestrian injuries have now become an important issue that has been on the agenda of several governments and public institutions for many years. The reasons for this change can be found, again, in economic factors – every injury brings a high social cost – but, also, human society underwent substantial changes during the last few decades. After more than a century of economic and technical growth, the capability of mankind to deeply influence the environment had become so high, and the available technology so powerful, that it became clear it is mandatory to keep them under control. In other words, it became clear that growth could not continue without limits, but rather, it had to face a new issue: the capacity to endure, that is, sustainability.

Sustainability is a complex and wide concept, which includes the capability of reducing as much as possible

the negative impact on the environment that the use of almost every new technology involves. Therefore sustainability deeply deals also with all transportation systems, because they pollute, they can be dangerous, they impact on people's everyday life, and, above all, they are essential.

Several studies on countermeasures aimed at reducing traffic injuries have been undertaken. Car manufacturers deeply increased the safety levels available inside cars. Back to the 1970s, objects like the steering wheel or the pedals, or, more generally, every part of a car's interior, could cause severe injuries in case of accidents, because they could hit the driver or passengers in an uncontrolled way. Car manufacturers then invested a considerable amount of money to reduce fatalities, and studied the problem. Nowadays, the steering wheel itself is not only able to avoid dangers in case of accidents, but also contains a safety system, the airbag, that sensibly reduces the consequences of a car accident.

Unfortunately, increasing road safety in the case of accidents involving pedestrians is sensibly more difficult. The EuroNCAP institution is involved in a number of projects related to this field (see [www.euroncap.com](http://www.euroncap.com) for an overview). Some of them aim at mitigating the consequences for a pedestrian being hit by a vehicle, for instance by modifying the shape of the car front, a sort of outer passive safety system. Similar ideas are under development since decades. Some patents regarding systems capable of mitigating collisions with a pedestrian date back to the beginning of the 1980s [2].

Following the idea of mitigating the impact energy, external airbags were also proposed. Several patents were submitted by the end of the 1990s [3, 4]. Such airbags should inflate in case the car hits a pedestrian, but still the problem of the collision sensor is open and hard to solve. Such sensors should detect collisions that are relatively soft, since the mass of a pedestrian is sensibly lower than that of a vehicle. Moreover, the sensor is required to be extremely fast and precise, because each false positive leads to the need of expensive service for restoring the external airbag system.

Together with passive safety systems, active systems were studied and developed in the last years. Since active safety systems should predict an imminent collision, they need sophisticated sensors and collision

detection systems. The task of preventing accidents from happening is extremely difficult, because it requires that the safety system understands the road scene at a high level, detecting potentially dangerous conditions. Such an understanding can be provided only by a computer system connected to sensors. The need for safety systems has therefore caused computer science to deeply enter into the automotive world, a process that had started with the very first electronic active safety systems, like ABS, few decades before.

Car manufacturers, together with public institutions, are still investing a huge amount of money in the study and development of such active systems. Several research groups are addressing the high number of problems that are related to road safety. Yet a system capable of a complete understanding of a generic road scene is not available, but technology is evolving rapidly, and now several systems exist that can analyze cars' or pedestrians' behavior over time, thus getting a high-level understanding of the road scene, even if limited to some aspects.

In order to achieve brilliant results, highly effective sensors should be used; possible selections include radars, laser scanners, cameras, or a combination of them. The number of solutions proposed so far is really huge, and substantial differences can be found in a number of design choices. This is clear even by just having a look at the various prototype vehicles that have been developed. Some of them are full of sensors, mounted in a fashion that is not realistic for an off-the-shelf car. Such vehicles are useful for experimenting with new solutions, understanding their possible advantages over more well-known system designs, without the effort of caring too much about the feasibility of the solution itself. On the other hand, vehicles designed with a high level of integration are vital to design systems that can be really installed onto a vehicle in the future. It often happens, in fact, that restrictions even on the hardware placement affect software algorithms dedicated to road scene understanding.

An example of the coexistence of these two kinds of design philosophy was clearly visible during the DARPA Urban Challenge [5], a competition for unmanned vehicles. Each vehicle had to complete a series of tasks, each one consisting in reaching a destination point driving inside an urban environment, dealing with traffic and intersections, and also

the other unmanned vehicles. Even if, in principle, no limitations were imposed on system design and number and type of sensors, vehicles that attended the race can be divided into the two categories described above.

Despite the high number of developed systems, as noted above, they are all exploiting, with few exceptions, the same types of sensor: radars, laser scanners, and cameras. Cameras can be further divided into three categories, depending on which part of the electromagnetic spectrum they are sensing: visible light, near infrared, or far infrared. They can be considered three different kinds of sensors since the data they provide have completely different characteristics, and usually undergo different processing.

Thanks to the high number of research groups working on road safety systems, a huge amount of algorithms has been developed for analyzing all kinds of data provided by the sensors described above. Also, due to the fact that those sensors are characterized by different and somehow complementary features, they are often used together. This led to the need for sensor fusion techniques, that is, techniques that, by exploiting measurements provided by more than one sensor, derive a single measurement. Such final result is more reliable than that provided by each sensor, as sensor fusion techniques know the feature of each sensor, and their reliability intervals.

As it has been described so far, computer vision applied to the automotive field is a really hot topic, and systems are being developed that successfully achieve a high and rather precise level of road scene understanding. However, beside system development, possible applications should be taken into consideration. The introduction of active safety systems based on such complex equipments into off-the-shelf vehicles is advancing at a rate that is sensibly slower compared with the advances in technology gained by research groups operating worldwide. Beside people acceptance of such systems, and legal issues, the requirements of the automotive world are extremely difficult to be met. If a system is capable of interfering with commands issued by a human driver, its acceptable fault rate will be extremely low, if not at no fault level. It is straightforward, on the other hand, that such a reliability level is really hard to be achieved, when dealing with such complex environment. The scenario is definitely challenging.

To understand how this kind of technology could develop, consider that still for a long time systems will not be able to reliably tackle the whole problem. Instead, some systems are already on the market, as equipment of luxury cars, which reliably perform simple and noninvasive tasks. As an example, some already available systems are meant to detect possible conflicts with overtaking vehicles when a lane change maneuver is undertaken by the driver. In this case the system issues a warning, like an acoustic signal. Some sporadic false positives do not have a strong impact; however, the driver will tend to ignore the system hints if failure rate is not extremely low. This kind of systems, that is, whose failure is not dangerous, but rather may be annoying, seems to be the next step on the roadmap of the introduction of active vision systems into the automotive world.

Given the previous considerations about the need to reduce accidents and the overview on future active safety systems, it is now worth analyzing into more detail which are the road scenarios causing a high number of fatalities or injuries. Several papers and reports exist on this topic; moreover, specific studies addressed the problem of accidents involving children [6]. This last problem is particularly alarming, since it involves people of very young age, whose perception of danger and understanding of road traffic rules is rather low due to their very young age. The danger comes from the fact that young children turn out to be often living in environments that are designed for adults. However, in the last decades modern societies started to invest in the safety of new generations, which is directly related to the sustainability of the societies themselves.

Describing the causes of pedestrian fatalities is, of course, an extremely complex task, and it also leads to subjective conclusions. For instance, the effect of car speed is debated. It seems to be a key factor in some studies, like [7]. On the other hand, in other studies related to children [6] it is stated that “higher vehicular speeds [...] were not associated with child pedestrian injury indicating that residential streets that had vehicles travelling at higher speeds were not at increased risk in the absence of other environmental risk factors.” Both studies point out that a main cause of accidents is the sudden appearance of the pedestrian in the road scene, as it often happens when a pedestrian comes



onto the street between two parked cars. This scenario is quite common, especially in residential areas, where, despite the low speed limits, a high number of pedestrian injuries happen. In most cases, the driver did not see the pedestrian before the accident occurred.

Investigating on main accident causes leads to a set of countermeasures that can be undertaken. In [7], three main actions are reported: “speed control, separation of pedestrians from vehicles, and measures that increase the visibility and conspicuity of pedestrians.” These involve road infrastructure, including some important details like the way cars are parked. The implementation of diagonal parking improves the visibility of pedestrians going toward the street, because it makes them walk in a diagonal direction, which is more easily seen from a driver’s perspective. This last example points out that protecting pedestrians does not necessarily mean the need for huge and expensive actions, but rather, that in some cases simple and well-thought expedients can be extremely effective.

Besides the modification of road infrastructure, improvement of pedestrian visibility can also be achieved by the active safety systems that were mentioned above. They seem to be the ultimate solution to the problem, because they are the only countermeasure to one of the major issues of the human drivers, that is, the tendency to get distracted, especially after driving for a long time. The activity of driving would require, in principle, a high and constant level of attention that a human driver can hardly provide. In particular, a driver usually cares about issues he is expecting, like a traffic light, but handles with much more difficulty a totally unexpected event, like a child darting into the street. For this kind of events, only the quick response of an active safety system can represent the final solution.

### Innovation of This Method

Instead of searching for pedestrians in a large area in front of the vehicle making no assumption on the external environment, the system presented in this work focuses on a specific urban scenario in which not only the detection of a pedestrian is of basic importance, but the danger of the situation can be clearly assessed as well.

In fact, in an advanced driving assistance system, a correct detection must be followed by the estimation of its position with respect to the vehicle and the environment.

As an example, a pedestrian detection system able to correctly localize all pedestrians present in the scene provides a huge amount of information that still needs to be filtered in order to be useful to either the driver or the onboard computer in charge of automatic maneuvers.

A possible filter may be implemented by fusing information coming from other systems, such as lane detection or other situation analysis engines, and CAN data. A pedestrian exactly in front of the vehicle may or may not be considered dangerous depending on the surrounding environment, as shown in Fig. 1.

This work is based on a totally different approach. Instead of detecting all possible candidates and assess their danger level, the system first analyzes the scenario, and then searches for possible pedestrians in specific positions for that particular scenario. This approach is called *scenario-driven search* or *SDS*. In this way, all detected pedestrians represent possible threats and no further filtering is needed (apart from a validation and a possible final tracking step).

The scenarios that are considered here refer to the most common urban situations in which the presence of a pedestrian causes a serious problem that could be mitigated by an early detection. In particular, when vehicles are moving on an urban road, the most common threat that a pedestrian may pose – therefore requiring a successful detection – is road crossing. Stopped vehicles on the road or on the road edges may occlude visibility, thus making the detection of the pedestrian more complex and requiring a real-time reaction.

The application of the *SDS* approach to this specific scenario is based on the localization of stopped vehicles followed by the search for pedestrians in their close proximity or in the areas partly hidden by them.

Stopped vehicles, whose edges trigger the search for pedestrians, may be parked cars on the road edge, vehicles temporarily stopped on the road, or vehicles queued in a line in front of a traffic light, zebra crossing, or simply jammed cars.

The first row of Fig. 2 shows some examples of situations in which the visibility of a crossing pedestrian is partly or completely occluded by stopped





**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 1**

When a pedestrian is localized but no reference with respect to the environment is provided like in (a), the detector is not able to assess the danger level. When environmental information are available, the very same pedestrian may become a threat (b), or may be in a safe position (c)

vehicles. The second row of Fig. 2 highlights, for each situation, the areas on which the system will perform a check for the presence of a possible pedestrian.

In other words, the approach is to focus on the detection of pedestrians appearing from behind occluding obstacles; pedestrians that are clearly visible in the frontal area of the vehicle need also to be detected. This function is also available on other systems [8, 9].

The idea of focusing on a specific scene or scenario (referring to a dynamic or a static environment, respectively) is not new to pedestrian detection systems. In 2002, Franke and Heinrich [10] developed a module able to detect balls (which are usually a strong signature of the presence of a child).

Another example of very specific systems is the one developed by Curio et al. [11] which was based on the visual localization of the specific moving pattern of human legs.

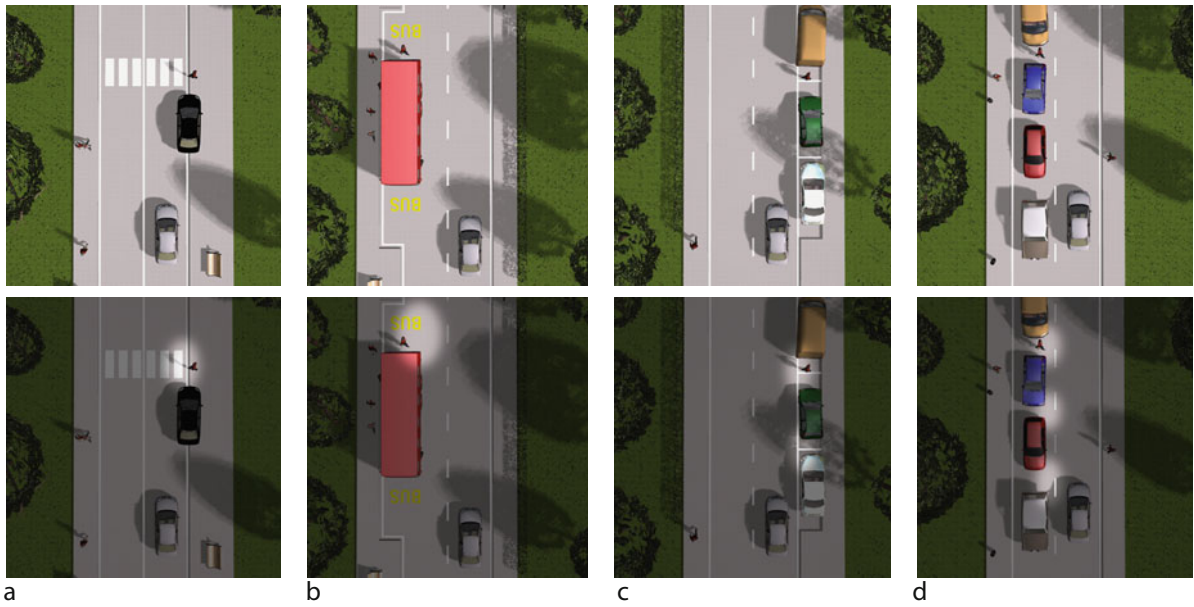
It is known [7] that parked vehicles, blocking the visibility of pedestrians, are one of the main causes of

accidents. Agran et al. [6] show that the number of parked vehicles along a street is the strongest risk factor for pedestrian injuries occurring in residential areas.

Although in these areas parking spaces should be arranged diagonally, there are situations, as shown in Fig. 2, in which vehicles stop temporarily on the road and their position cannot be predetermined and carefully controlled as in the case of parking lots. Although some of the situations of Fig. 2 refer to specific urban areas (i.e., zebra crossings and bus stops) that could be specifically enhanced by intelligent infrastructures aimed at warning oncoming vehicles, other situations can happen in any part of the road network, making the installation of specific warning infrastructures impractical.

The main characteristics required by a system like this are:

- To detect pedestrians quickly, given the short sensing range and the particularly high risk of a collision



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 2**

Situations considered in this work (*first row*) and areas of interest considered for the detection of a possible pedestrian (*second row*). (a) A crossing pedestrian is hidden by a parked vehicle. (b) A pedestrian is crossing the road behind a stopped bus. (c) A pedestrian is appearing between two parked vehicles. (d) A pedestrian is crossing the road between two vehicles stopped on the other side of the road

with a pedestrian suddenly appearing behind an obstacle

- To detect pedestrians as soon as they appear, therefore even when they are still partly occluded
- To limit the search to specific areas, determined by a quick preprocessing

Figure 3 shows the coverage of different sensing technologies (laser and vision) in a specific scenario considered in this work.

The fusion of a laser scanner and vision can provide a quick and robust detection in case of suddenly appearing pedestrians. The laser scanner provides a list of areas in which a pedestrian may appear, while the camera is able to detect the pedestrian even when he/she is not yet visible to the laser scanner.

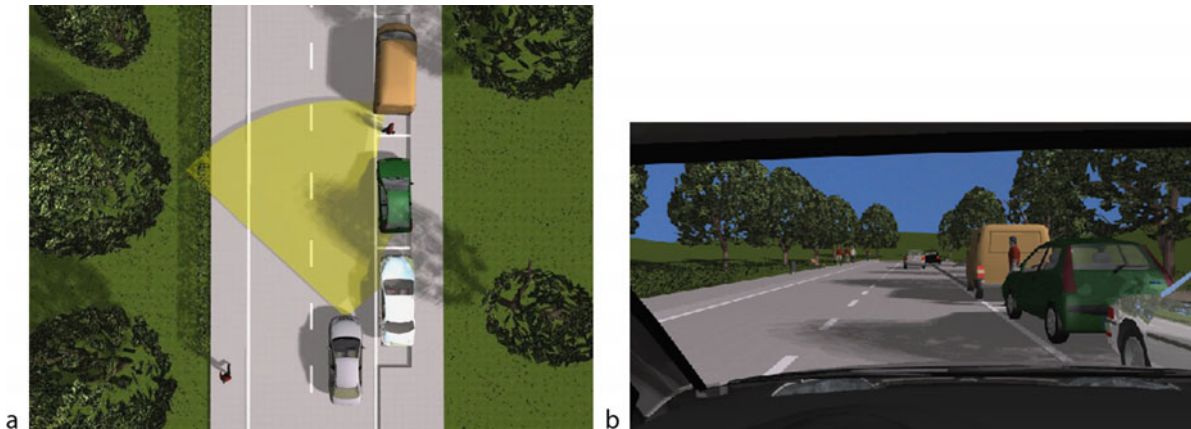
A remarkable improvement of road safety can be obtained thanks to the use of both reversible and nonreversible driving assistance systems. Three levels of intervention are defined:

- Once the pedestrian is detected with a sufficiently high confidence level, a warning is sent to the driver.

- Should the driver not react promptly to the warning, the system would issue a second level of warning by blowing the vehicle's horn. This second warning is still considered a reversible system, although it is much more invasive than the former. The aim of this loud warning is to attract the attention of both the pedestrian itself and – once again – the driver.
- In case the danger level is not reduced thanks to a prompt reaction of the driver (or the pedestrian), the intelligent vehicle will jump into the third level and trigger a nonreversible system. In this case it is automatic braking.

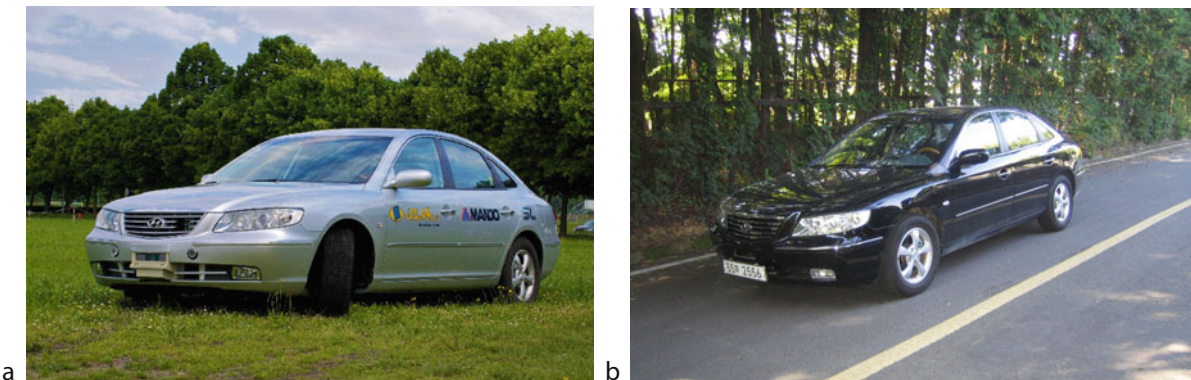
This being a nonreversible and very invasive system, its triggering must be preceded by an extremely careful analysis of the danger level and requires the complete processing to be thoroughly tested with respect to false detections. In the case of nonreversible systems, the number of false positives is even more important than the number of correct detections.

In the literature [12] many different pedestrian detection systems have been surveyed, and a large



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 3**

A pedestrian partially hidden by parked cars may not be detected by a laser scanner positioned in the front bumper but can be detected using vision, even if partially occluded



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 4**

The Hyundai Grandeur test vehicle equipped in Parma, Italy (a), and the replica in Seoul, Korea (b)

number of test methods for pedestrian detection systems have been proposed, which are based on single frame analysis, temporal analysis of continuous video streams, or event based.

Their performance is still far from the ideal system; this is the reason why the approach proposed in this paper is regarded as an alternative to conventional systems. Pedestrians are searched in situations of clear danger by limiting the search to specific areas. Besides being quicker than other systems, this method aims also at reducing the number of false detections to zero.

The system presented in this work has been developed and tested by VisLab in Parma on a Hyundai

Grandeur prototype vehicle; Mando is using an exact replica of this system in Korea (see Fig. 4) to double testing time.

### State of the Art

As long-time pursued advanced driver assistance systems (ADAS) and active safety vehicles (ASV) such as lane departure warning (LDW), blind spot detection (BSD), adaptive cruise control (ACC), precrash safety (PCS), and collision warning/avoidance (CW/CA) become commercialized [13–15], interest of research groups and transportation authorities moves to newly

emerging technologies or candidate projects such as intersection accident prevention [16], intelligent parking assist system [17], and active pedestrian protection system (APPS). Particularly, APPS attracts a great interest because it is directly related with a vulnerable road user who holds a significant portion of traffic accident fatalities. The success of ACC, PCS, and CW/CA does not guarantee the success of APPS because the pedestrian is relatively small and their appearance is various and continuously changing. However, they are expected to provide abundant elementary technologies and economic backgrounds. In particular, progress of electronics and sensors, such as high performance embedded processor, monolithic microwave-integrated circuit (MMIC)-based compact radar, multilayer scanning laser radar, and high dynamic range CMOS (HDRC) camera, are regarded as the enabling technologies for APPS.

Recently, Gandhi and Trivedi provided a great survey of pedestrian protection systems [12]. They classified pedestrian protection approaches into four categories: infrastructure design enhancement, passive safety system involving vehicle design, active safety system based on pedestrian detection, and collision prediction. The pedestrian protection system designed to be installed on board can be divided into passive and active approaches according to whether the operation time is before or after vehicle–pedestrian collision. If a system is designed to prevent the collision or mitigate its damage, it is categorized as an active approach. Contrarily, if a system is designed to minimize the pedestrian's damage after the collision, it is categorized as a passive approach. The collision prediction could be differentiated from the active safety system in the point that the sensing devices are installed on infrastructure rather than on board.

Infrastructure enhancement to reduce pedestrian-related accidents is supposed to be the most traditional approach. The approach can be divided into three categories of countermeasures: speed control, pedestrian vehicle separation, and measures to increase visibility and conspicuity of the pedestrian [12].

The first and second could be easily understood by investigating traffic calming which is a kind of movement or concept [18]. Definitions of traffic calming vary, but they all share the goal of reducing vehicle speeds, improving safety, and enhancing the quality of

life. Traffic-calming measures can be separated into two groups based on the main impact intended. Volume control measures are primarily used to address cut-through traffic problems by blocking certain movements, thereby diverting traffic to streets better able to handle it. Volume control measures also include divertive or restrictive measures such as full closures, half closures, diagonal diverters, and median barriers. Speed control measures are primarily used to address speeding problems by changing vertical alignment, changing horizontal alignment, or narrowing the roadway. Vertical deflection includes speed humps, speed tables, raised crosswalks, raised intersections, and textured pavements. Horizontal deflection includes traffic circles, roundabouts, chicanes, and realigned intersections. Horizontal narrowing includes neckdown, center-island narrowing, and chokers. US DoT (Department of Transportation) FHWA (Federal Highway Administration) reports [19, 20] show that well-designed and implemented traffic-calming techniques can have a number of beneficial impacts for bicyclists and pedestrians. The reduced vehicle speeds associated with such projects can reduce both the severity and incidence of motor vehicle/bicycle/pedestrian crashes and can make bicyclists and pedestrians feel more comfortable in traffic.

Pedestrian visibility can be increased by improving roadway lighting since a majority of pedestrian fatalities occur at nighttime. Since parked vehicles block the vision of drivers as well as pedestrians, removing on-street parking and implementing diagonal parking in residential streets would help in reducing accidents especially involving children [12].

According to the analysis of vehicle impact zones in pedestrian–vehicle collisions, vehicle front is responsible for most pedestrian injuries: 71.1% according to US national highway traffic safety administration (NHTSA) data. Almost 80% of the fatalities occur when the pedestrian is struck by vehicle front, whereas most of the remaining fatalities occur when the pedestrian is hurled to the street [21]. Accident analysis also proved the importance of head injury [22, 23]. Although upper and lower extremities, as well as head, are the most frequently injured body regions in passenger car-to-pedestrian accidents, 62% of the pedestrian fatalities were caused by head injuries. Pedestrian head suffers both a direct impact with stiff



vehicle structures and a large degree of rotational accelerations [24]. In particular, stiff structure of windshield including A-pillar should not be overlooked. An analysis of all vehicle parts which caused head injuries shows that impact on the windshield causes 44.3% of head injuries. Impact on the windshield causes twice as often head injuries as at the bonnet, and three times often for severe head injuries. Head impact on the A-pillar causes severe injuries in 82% of the cases. Other vehicle parts mostly cause only slight injuries [24].

Therefore, passive safety measures are being developed by installing energy-absorbing devices in the vehicle front structures, on which pedestrian head hits, such as hood (bonnet), windshield lower part, and A-pillar. A passive pedestrian protection system includes an impact-reducing front structure, active bumper, active hood (hood lifting, active bonnet), and pedestrian airbag. Active bumper is designed to provide the appropriate stiffness to cushion the impact while at the same time providing support of all parts of the limb to limit knee joint lateral bending [25]. Cushioning method of active bumper includes foam energy absorber, molded plastic energy absorbers, air-filled energy absorbers, flexible plastic beam, bumper deploying, crush cans, outside additional structure, foam-encapsulated metal, and steel energy absorber [25]. Active hood system detects vehicle-pedestrian collision using bumper-installed pedestrian crash sensor and acceleration sensor, and then two actuators lift the rear part of the bonnet approximately 100 mm within 60–70 ms after the leg-to-bumper impact [26]. Pedestrian protection airbag deploys an airbag through a gap between windshield lower part and bonnet lifted by the active hood system [27]. The airbag is deployed in U-character shape and covers the windshield lower part and A-pillar lower part. When up-lifting bonnet and airbag are used simultaneously, the impact absorbing effect is significantly larger than when only one measure is used. It is because vehicle front regions critical to pedestrian head impact cannot be effectively covered with only one measure [24]. Intelligent Vehicle Safety System (IVSS) seems to be a good example: Active bumper, active hood, and pedestrian protection airbag is developed simultaneously and they are expected to operate as a suite of tools [27]. Autoliv also developed and provided active hood and pedestrian protection airbag as a pedestrian protection tool chain [28, 29].

Such a development of passive safety measure is supposed to have been significantly influenced and encouraged by the reinforced requirement of pedestrian-vehicle collision. The European Enhanced Vehicle Safety Committee (EEVC) Working Group (WG) 10 and WG 17 have developed test procedures to assess the level of pedestrian protection for vehicle fronts. Based on the EEVC WG 17 report, legal requirements have been derived. The European directive (2003/102/EC) consists of head impact, upper leg impact, and lower leg impact. The requirement was scheduled to be enforced in two phases, that is, 2005 and 2010 [30]. The EU test levels are 1,000 head-injury criteria (HIC) for 2/3 of the hood surface and 2,000 HIC for the remaining 1/3. HIC is an internationally accepted acceleration-based measurement for violence against the head. HIC values under 1,000 imply that the risk of life-threatening injury is 15% or less. However, the curve rises sharply; at 2,000 HIC, the fatality risk is almost 90% [28].

Active safety systems are based on pedestrian detection and collision avoidance measures. Considerable research is being conducted by various groups for designing pedestrian detection systems. Such systems can employ various types of sensors and computer vision algorithms in order to detect pedestrians and predict the possibility of collisions. The output of the systems can be used to generate appropriate warnings for the driver or to perform autonomous braking or maneuvering in case of imminent collision [12]. Naturally, if impending collision could not be avoided, the fact can be sent to the passive pedestrian protection measures and precrash safety system to help their preparations. Even when the autonomous braking cannot prevent collision, it can mitigate the pedestrian's damage. The reduction of collision speed leads to a reduction of the induced head accelerations and results in a shift of the relevant head impact regions away from stiff windshield lower parts. The core technology of active pedestrian protection system is pedestrian detection because the active braking function is already implemented by several active safety systems such as PCS, CW/CA, and full-range ACC [13, 14].

Pedestrian detection can be divided into five categories according to the used sensors: thermal infrared camera-based, visible camera-based including near infrared camera, millimeter wave radar-based,

communication-based, and sensor fusion-based. Even though scanning laser radar is frequently used by pedestrian detection system, it is not considered in a separate category because it is not used as a standalone sensor but as a participator in sensor fusion with visible camera.

Thermal infrared camera-based is specialized for nighttime pedestrian detection, for example, night vision, and is based on the point that the human body temperature is higher than ambient temperature. Although it can easily detect pedestrians even in dark illumination conditions, it loses the advantage in daytime and its cost is comparatively high for all grades of vehicle.

David Gernóimo et al. provide a great survey of visible camera-based pedestrian detection [31]. They propose a general architecture of a visible camera-based pedestrian detection system consisting of six phases: preprocessing, foreground segmentation, object classification, verification, tracking, and application. Preprocessing, foreground segmentation, and appearance-based object classification are very similar to vehicle detection case [32]. Foreground segmentation for candidate generation can be performed using knowledge-based, stereo-based or motion-based methods. Appearance-based object classification uses generally proven learning algorithms such as support vector machine (SVM), AdaBoost, and neural networks. Characteristic approaches are contour tree-based template matching devised by D. Gavrila and part detection-based to cope with the pedestrian's continuously changing contour and appearance. What kind of feature and classifier is the best for pedestrian detection is a problem under fierce discussion. To answer the question, Mohamed Hussein et al. compared performances of two promising features, that is, histograms of oriented gradients (HOG) and region covariance (COV), while the total system was implemented by a rejection cascade of boosted feature regions [33]. Markus Enzweiler and Dariu M. Gavrila compared four state-of-the-art pedestrian detection systems: wavelet-based AdaBoost cascade, linear SVM with HOG, neural network (NN) with local receptive field (LRF), and hierarchical shape matching combined with texture-based NN/LRF classification [34].

Prof. Hermann Rohling's team of Hamburg University of Technology developed distinctive

pedestrian detection systems using only automotive millimeter wave radar [35, 36]. They discovered a fact that radar echo signals from a walking pedestrian contain characteristic information that differs from other backscattering objects like cars and trucks. Automotive radar sensor can measure not only target range with high resolution but also velocity with an accuracy of a few centimeters per second. When the location and velocity of target are measured, the pedestrian has a narrow location range and a wide velocity range because of its moving arms and legs while the vehicle and pillar have a narrow velocity range with a wide and a narrow location range, respectively. Furthermore, the tracked velocity of pedestrian shows regular oscillation with a certain low frequency, which is a strong indicator and a signal feature to distinguish between pedestrian and car. Using features extracted from radar signal, the pedestrian can be classified from moving vehicle and stationary objects.

Communication-based pedestrian detection is the newest approach. It is derived from the idea that radio-based communication does not need line of sight and could solve hidden pedestrian problems. Although the feasibility of mobile communications such as global system for mobile communication (GSM) and universal mobile telecommunications system (UMTS) was studied, they have many disadvantages. Every node should have a precise global positioning system (GPS) and peer-to-peer communication should be added to meet latency requirement [37]. EU project WATCH-OVER and German project AMULETT investigate the possibility of short-range communication for pedestrian protection [38, 39]. In this approach, vehicles transmit interrogations and transponders embedded into vulnerable user's accessories such as school bag, bicycle, and motor bicycle return identification messages. With the multi-antenna system, the vehicles can recognize the directions and distances of the transponders. Furthermore, the identification message can provide a user's information such as agility, age, and moving pattern. Candidate short communication technologies include IEEE 802.15.4 (Zig-Bee), radio frequency identification (RFID), and IEEE 802.15.3a ultra wide band (UWB) radio. Even if this last approach is very interesting, it is based on the strong assumption that all road users have transponders: at present this is unrealistic.



## Setup

This section describes the setup, starting from the perception technologies, going through the sensor suite and actuators, and concluding with the processing system.

Being designed to address an urban scenario in which the prototype vehicle is running close to stopped vehicles, a limitation on vehicle speed and detection range can be accepted. Low-to-medium vehicle speeds up to 50 km/h and a detection range of about 40 m can be considered as a safe choice.

One of the constraints of this problem is the detection of stopped vehicles and other standing obstacles to be detected; a laser-based solution sufficiently fit this requirement robustly enough supplying both distance measurement and shape classification. The speed information is commonly missing in laser data. However, some laser scanners, like the one selected for this application, can provide interlaced data. For these devices, a number of interlaced scans, taken at different times, are grouped together in a single, high-resolution scan. Obstacles speed can be estimated by analyzing these interlaced data together with vehicle inertial information. Using obstacle speed a classification between standing and moving obstacles is straightforward.

Vision, is the key technology to detect pedestrians even if they are partially occluded. Since distance measurements are already available from the laser and no further 3D reconstruction is required, a single NIR camera satisfies the requirements for both daylight and night scenario.

The camera selected for this system is an AVT Guppy F-036B. Its sensor of  $752 \times 480$  pixels has a wide aspect ratio which is beneficial for automotive applications since relevant information is often contained in lateral area. The sensitivity covers both the visible and the NIR spectra. During the night, object detection is improved combining the high response in the NIR domain using and specific illumination. Additional headlamps are mounted in front of the vehicle. A NIR LED headlight with  $25^\circ$  of aperture is mounted in front of the radiator, while the headlamp blocks customized by SL can illuminate in the visible domain, using a bifunctional lamp (low-beams and high-beams), and in the NIR domain using additional specific lamp.

The laser scanner is a SICK LMS 211–30206. This laser has a horizontal scanning angle of  $100^\circ$  with angular resolution selectable from a minimum of  $0.25^\circ$ ; the detection range goes up to 80 m; a fog correction feature is also available.

All these characteristics well fit the application requirements, as a lot of commercially available lasers. But this one also features interlacing. Selecting  $1^\circ$  of resolution and a phase-gap of  $0.25^\circ$ , four subsequent rotations are included in the same scan. Four rotations are performed in 53.28 ms, thus every rotation takes 13.32 ms. This time is comparable with that of the events occurring in a driving scenario. This means the scans deformed mainly depending on the vehicle movement. This effect becomes more evident using the interlacing feature.

Since the camera and the laser scanner are not synchronized, there is a continuously variable time shift between the two captured sample streams; however the displacement due to the non-synchronization is considered negligible due to ratio between the vehicle speed, which is low, and the processing rate, which is high.

The laser scanner and NIR headlamps are integrated in the front bumper, as shown in Fig. 5. The NIR camera is placed inside the driving cabin near the rearview mirror as shown in Fig. 6.

An industrial grade Mini-ITX PC running an Intel Core 2 Duo-based CPU is fixed in the boot. The camera is connected through a FireWire A controller integrated on the motherboard, while an external RS422 to USB adapter is used to connect laser. Inertial data



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 5**

A detail of the front bumper showing the laser scanner integration and the headlights

produced by off-the-shelf sensors are gathered through the CAN bus using a USB adapter.

On the actuator side, the braking system installed on the test vehicle was replaced with Mando's MGH-40 ESC plus. The braking strength can be controlled by sending appropriate CAN messages to this ECU. In this unit is also available a deceleration control interface (DCI) used for high-level functions, such as adaptive cruise control (ACC) and precrash safety (PCS). The DCI may receive multiple deceleration commands. The



Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 6

How the camera is installed inside the cabin

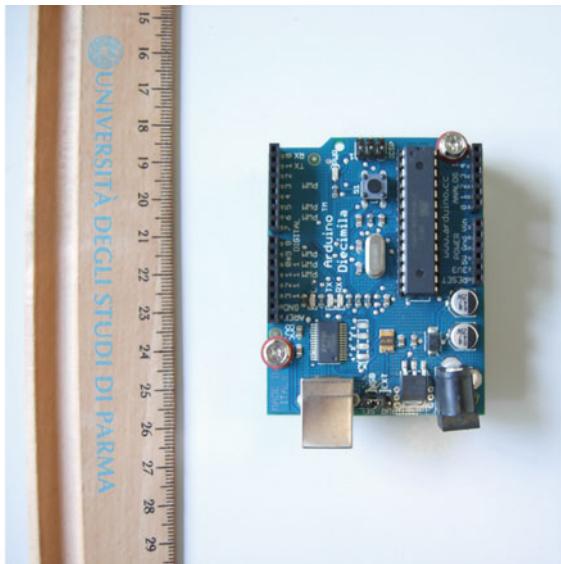
desired deceleration is selected depending on priorities and vehicle status. A feed-forward controller uses the deceleration output as control signal in a feedback loop to physically operate on wheel brake pressures. The braking strength can be controlled in the range  $[0, 1.0]$  g with a resolution of 0.01 g, a stable error lower than 0.05 g, and response time of 0.3 s. Deceleration is controlled by the DCI in cooperation with other active control system Electronic Brake Distribution (EBD), Anti-lock Brake System (ABS), and Electronic Stability Control (ESC).

The horn has been modified to receive a control signal from both the manual command and a USB I/O board (see Fig. 7a) connected to the processing unit and detected as a serial port.

A new 100 Hz yaw rate sensor, manufactured by Siemens VDO and Mando, is connected through the processing system via CAN bus. This sensor provides yaw rate with an accuracy of  $0.0625^\circ/\text{s}$  and also lateral acceleration. Figure 7b shows the sensor.

### Focus of Attention

Two methods for laser data classifications are developed to improve algorithm robustness. Pulses are first clustered into straight lines for both methods; then the



a



b

Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 7

(a) The USB I/O board used to control the horn, and (b) the special yaw rate sensor

first one classifies obstacles only considering the last scan, while the second one verifies the obstacle position and speed using subsequent scans.

Significant shifts between subsequent laser scanner measurements affect the measured obstacle shape as a distortion, and cause clustering or classification problems.

Vehicle movement in terms of rotation and translation is measured using the ESP CAN box. Having this information available, each pulse in the scan can be corrected removing the shape distortion.

Translation is corrected using vehicle speed, while using the steering wheel angle to compute the yaw rate allows to completely specify the rotation correction matrix.

Yaw rate can be computed as follows:

$$\text{Yawrate} = (V_{fr} + V_{fl}) / (t_w \cos \delta)$$

where  $V_{fr}$  and  $V_{fl}$  are front right and front left wheels speeds, respectively,  $\delta$  indicates the wheel angle, and  $t_w$  is the vehicle front track.

The additional yaw rate sensor installed on the vehicle is used to measure directly this value, since indirect computation is highly affected by noise.

Figure 8 shows raw and corrected laser data referring to nonmoving obstacles. The four laser rotations that create a scan are clearly visible before the correction, while the obstacle appears as a single line after the correction.

After data correction, moving obstacles captured from interlaced scans are characterized by four parallel lines, as shown in Fig. 9. Obstacle classification is performed relying on this information.

Pulses belonging to the same rotation are connected together as a first step forming chains of segments or cluster. Points which are not joinable to any chain of the same rotation are checked for merging with points of other rotations considering proximity only. Points not connected to other points or close to the limit of the laser scanner range (about 80 m) are permanently discarded.

After clustering, moving obstacles are classified by identifying four different and parallel clusters, while standing obstacles are detected when the clusters are overlapped.

Up to this step, chains of segments are created by joining pulses without any additional information.



Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 8

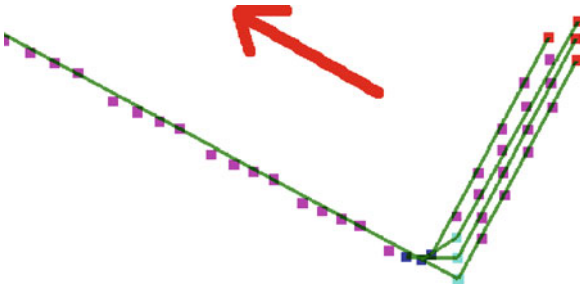
Laser data: (a) without correction and (b) with correction



Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 9

After data correction a moving obstacle is still represented by four different lines

When adjacent segments have approximately the same orientation, they are merged into a longer segment, preserving the obstacle shape, but reducing the amount of information supplied by the sensor. Straight lines connecting the chain start and end points are obtained merging a chain of adjacent segments.



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 10**

A moving vehicle and its motion direction: the four *parallel lines* correspond to its rear bumper, while the *single line* corresponds to its left side. *Blue points* are polylines start points, *cyan points* are corner points, *red points* are polylines end points, and *violet points* are generic line points

The distances between a line and all its internal points are computed and, if the maximum between these distances is larger than a threshold, the line is split into two separate lines.

These steps are iterated while the maximum distance is larger than a threshold. The result is that each chain is therefore finally segmented into a polyline.

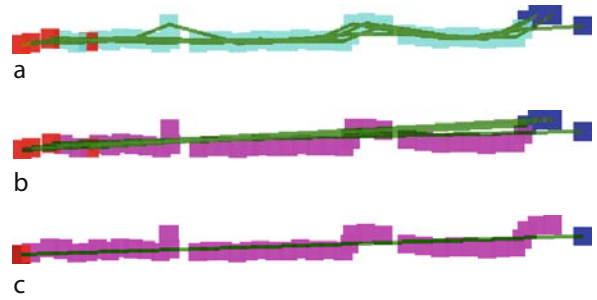
Figure 10 shows a moving vehicle: The rear bumper is framed as four parallel lines, while its side – parallel to the vehicle movement direction – is marked by a single line.

After data correction, points clustering and segment merging, each obstacle detected in the laser scanner field of view, is identified by four lines, one for each laser rotation that composes the whole high-resolution scan. Static and moving obstacles can be located and classified in a robust way using the following observation. A static object has four lines that are overlapped on a single shape, while a moving obstacle has the lines parallel but not overlapped. Figure 11 shows the algorithm steps.

Although this process is very simple and effective, when vehicle pitch is significant, false negatives are possible.

Polylines obtained so far can only be classified thanks to their size and shape. Obstacles can be divided in four categories:

- Possible pedestrian
- Road infrastructure



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 11**

The steps of the algorithm: (a) data clustering, (b) approximation using polylines, and (c) lines merging: all the points are merged into a single line; in order to simplify the following steps, small drifts are ignored

- L-shaped obstacle
- Generic obstacle

Obstacles with limited size are classified as pedestrians, while obstacles almost parallel to the vehicle and with a large size are classified as road edges (guard-rails, buildings, road infrastructures, etc.).

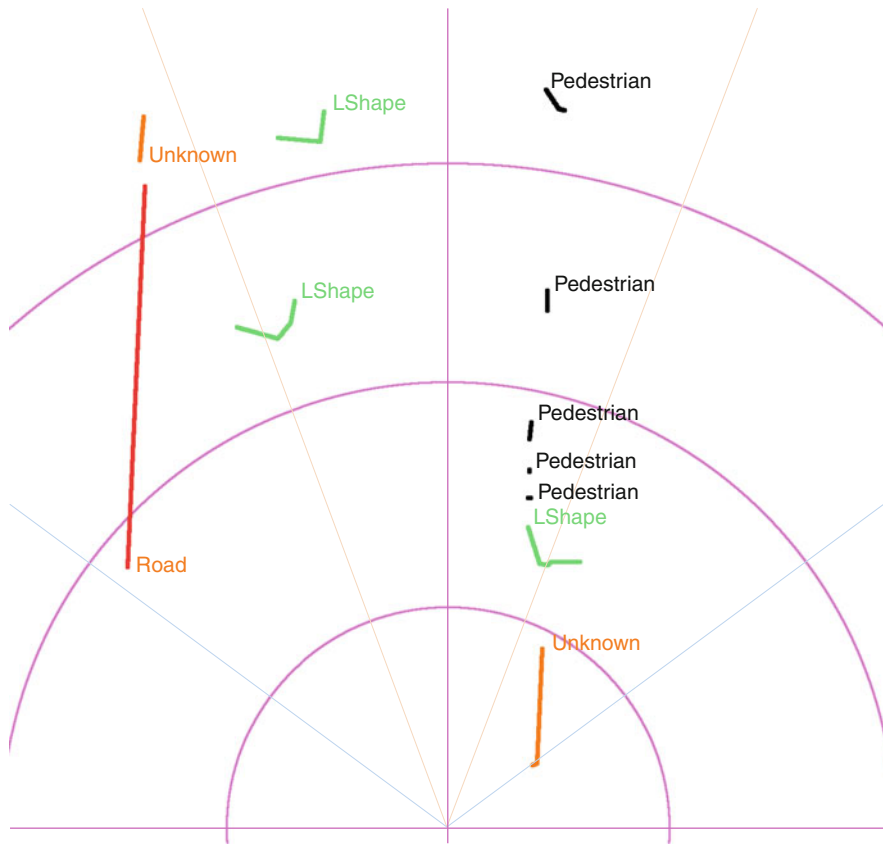
L-shaped obstacle patterns are detected using a simple and fast method, based on line orientations. All obstacles not yet classified are tagged as generic obstacles; Fig. 12 shows an example.

Information about obstacle movement – already estimated in the previous step – is also stored for further usage.

The procedure described in the previous subsection classifies all small obstacles as possible pedestrians. However a number of false positives may indeed be present under unfavorable conditions.

Parked vehicles and other fixed obstacles along the road are used to identify critical areas in front of the vehicle, to focus the detection in the immediate proximity of these areas where pedestrians can suddenly appear and become dangerous.

Ego-motion information is used to align and rotate polylines provided by previous scans; a check on the overlapping between polylines identified in current and in previous frames is then performed to assign a category to the object between: moving obstacle, static obstacle, changing shape obstacle, and new obstacle.



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 12**

Obstacle classification: all possible classifications based on shape are represented in the image

Moving obstacles have small spatial overlapping if the time window is large.

If no correspondence is found between the current polyline and those from the past, the object is classified as a *new obstacle*.

The areas of interest in which vision will search for pedestrians is determined using static obstacles; obstacles whose shape is changing over time are also of basic importance. They may contain a pedestrian in a smaller region of their shape.

The centroid of each polyline is computed to track obstacles over time. Centroids are used as representative points to simplify the computation. A description of a more complex laser scanner data processing based on comparison between each point of polylines can be found in [40]. A different classification is also made using distances between the centroid of the last and previous scans.

Polylines are labeled using the information contained in the distances between the current centroid and all the past ones as follows:

- Fixed obstacle: if all distances are below a certain threshold. This threshold, to detect, is increased according to the distance of the obstacle (to compensate for a decreasing accuracy) and the time gap (to partly compensate for errors in ego-motion reconstruction).
- Moving obstacle: if the distances grow over time.
- Unknown obstacle: this label is given if the object can neither be labeled as fixed nor moving. This may happen the first time an obstacle appears or when the detection is not stable due to bad reflections of the lasers beam.

The three previous classification results are considered to compensate for laser data processing



inaccurate results. The polyline is classified with the most frequent one.

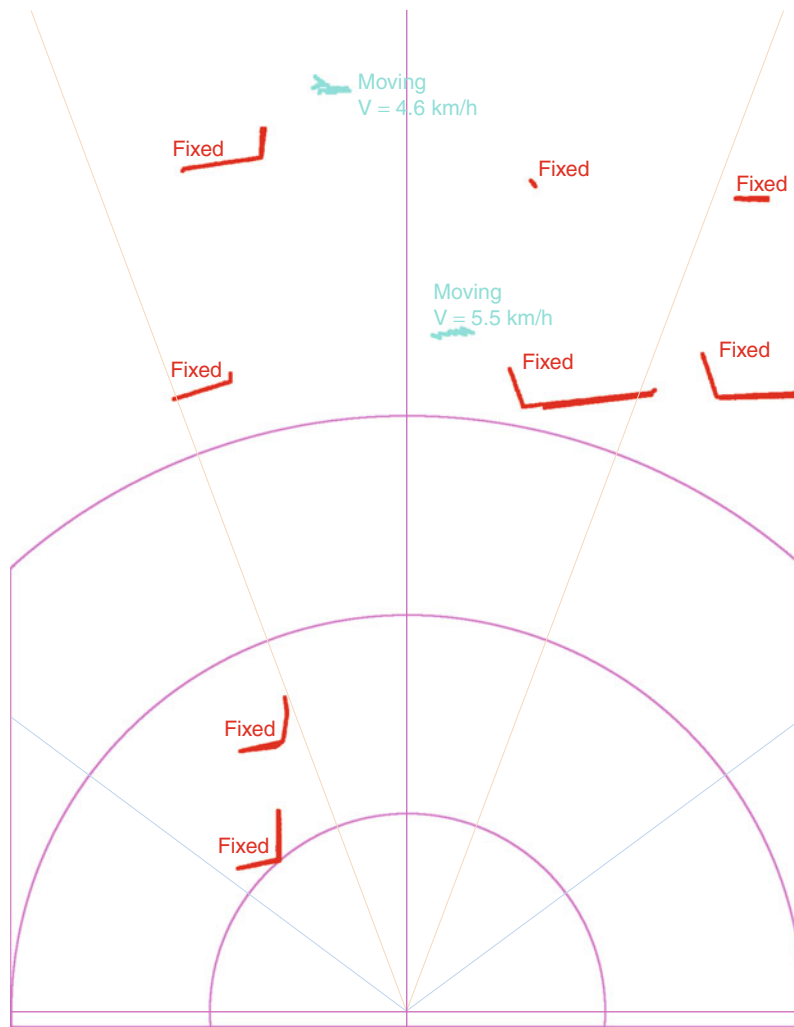
Example results are shown in Fig. 13 with an estimation of the object speed.

The following steps of the algorithm use the processing results on the labeled polylines. This information is used to sense the static (obstacles and vehicles) position and moving (pedestrians and vehicles) obstacles in the environment. When obstacles are identified as static by the second classification stage, they are used to define the environment structure. Driving corridor is defined as the area that may be reached by the

vehicle in the near future. Fixed obstacles are used to build the driving corridor, namely, as shown in Fig. 14.

The driving corridor may have multiple paths or branches, each with a minimum width larger than the vehicle one; branches with a width narrower than this threshold are discarded as closed. The corridor does not necessarily correspond to the free space in front of the vehicle, since it is built considering fixed obstacles only, and may also include moving obstacles.

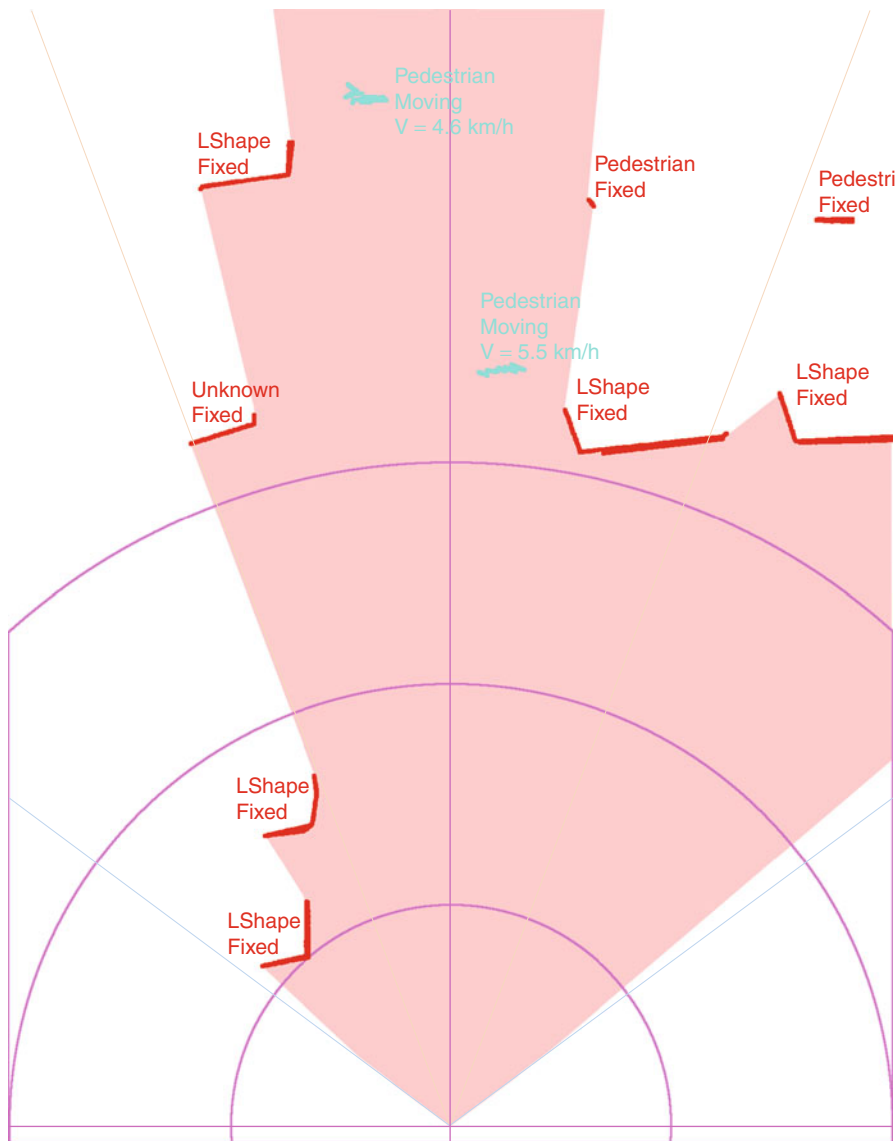
This algorithm is focused on the detection of suddenly appearing pedestrians; the detection is concentrated on areas hidden by a static obstacle or



Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 13

Obstacle classification: all possible classifications based on movement are shown. Note that moving polylines change their position in world coordinates





**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 14**

The pink zone represents the driving corridor. Obstacles moving inside the corridor, or near its edges, with a speed and a size typical of a pedestrian, are tagged as possible pedestrians

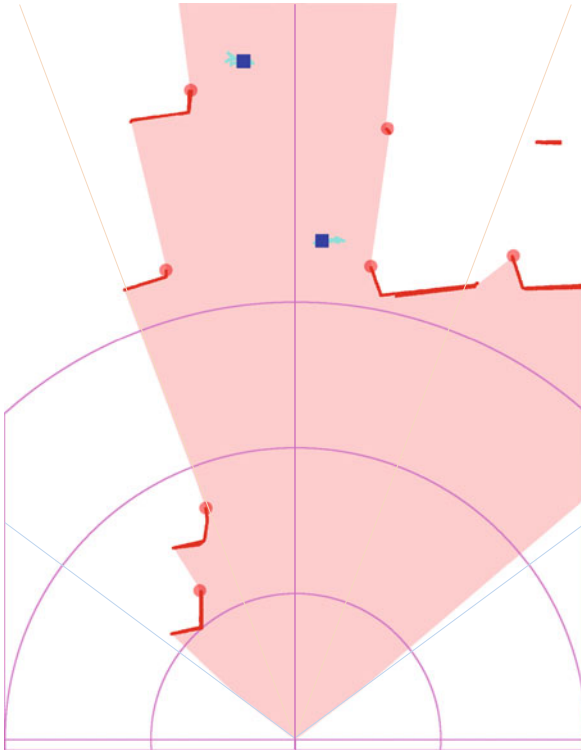
between two static obstacles. These areas are usually located along the corridor edges. Static obstacles are used to build the corridor, and also to identify dangerous areas, that is, areas in which pedestrians may appear. Dangerous areas are located behind the furthest point of each static polyline as shown in Fig. 15.

Obstacles moving inside the corridor, or near its edges, with a speed and a size typical of a pedestrian, are tagged as possible pedestrians.

### Vision Fusion

The regions of interest are 2D areas in world coordinates; their corresponding areas in the image are then located thanks to camera calibration and perspective mapping.

Once the areas of attention are located, the search for pedestrians is triggered in these areas. Specific image windows are defined using a perspective



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 15**

Dangerous areas and possible pedestrians: red circles represent dangerous areas, while blue boxes identify possible pedestrians

mapping transformation, considering 90 cm as pedestrian width and 180 cm as pedestrian height. These image areas are resampled to a fixed size ( $24 \times 48$  pixels).

In order to reach good classification during both day and night, two different systems are developed. The idea is to use a simple and specific classifier for nighttime images, in which pedestrians' shapes clearly appear, and to use more complex classifiers (based on cascade methods) for the day.

The nighttime algorithm is based on the assumption that, thanks to NIR illuminators, pedestrians and other standing objects represent the brightest areas in the image with the exception of light sources. Therefore, after removing light sources, it is possible to detect obstacles with a binarization process and classify them according to their shape.

At the end of the preprocessing steps, a binarized image is obtained. The standing objects (including pedestrians) are remapped as white, while the background is remapped as black. As the NIR headlamps illuminate more the legs of the pedestrian, these first phases are operated separately for the upper and lower part of the image.

After these operations, the procedure continues to check for pedestrian trunks and legs by respectively analyzing the upper and lower half of the image. First of all open legs are searched for; if open legs are not found, an algorithm that checks for closed legs is applied, and, finally a trunk presence check is performed.

The open legs search is based on striped skeletonization. For each row of the image, the white areas present are narrowed to a single point. In such a way it is possible to remap legs as lines. Obtained lines can be transformed to straight lines, and the position of these lines can be finally studied to identify legs.

Search for trunk and closed legs are very similar, differing only in some details and thresholds. The captured images contain a very low level of detail and it is often inadequate for a complex analysis. For this reason a simple procedure is implemented that tries to make the most of the few details that characterize the pedestrian. The algorithm searches for white areas and considers their aspect ratio to confirm the presence of trunk or closed legs.

At the end of the algorithm four rates are assigned, which range from 0 to 1. The votes are:

- Vote of the trunk according to the correspondence with a model and to dimension.
- Vote of the legs according to the correspondence with a model and to dimension.
- Open legs rate which is obtained with the ratio between feet position distance and maximum feet distance.
- Pedestrian vote which is obtained making an average of the body and legs vote. The vote of the legs has more weight if they are open because it is a dominant characteristic of a pedestrian.

All these ratings will be analyzed at a later stage of the application to properly determinate if the figure is recognized as a pedestrian.

AdaBoost (a short for Adaptive Boosting) is a machine-learning meta-algorithm. It is used to select the combination of a weak classifier to obtain a precise classification.

Using an appropriate training set it is possible to compute the correct weight to be set to each classifier and obtain a robust classification. AdaBoost is a technique widely used for the classification of pedestrians [41]. Haar features are firstly chosen for the weak classifier [42]. Different Haar features are selected for each iteration, as suggested by Viola and Jones [43]. Haar features are very simple features that can be used, together with a meta-algorithm similar to AdaBoost, to classify a lot of objects. They consist of a very simple pattern that can be moved and stretched on the image to be compared to image integral, in order to obtain an accurate classification.

As Haar features are very generic, the idea is to develop other features, specific for pedestrians, in order to reach a better classification using AdaBoost again. These features are based on a typical pedestrian shape and body parts aspect. To detect a pedestrian, the algorithm provides a vote to each feature of the following list:

- Open legs shape
- Left step legs, little mask
- Right step legs, little mask
- Left step legs, big mask
- Right step legs, big mask
- Head top, little mask
- Head top, big mask
- Head, with circle formula
- Pedestrian's M function
- Horizontal border function
- Tree and pole search function

To increase the potential and the recognition performance, an additional stage, called AdaBoost cascade, is developed.

In this stage the votes from the classifier implemented with Haar features, with ad hoc ones, and nighttime classifier are merged and combined to obtain an additional level of classification. The votes provided by the previously described algorithms are used as input to make a second kind of AdaBoost classifier.

As the number of outputs from previous classifiers is very low, and some tests pointed out that they cannot

be used to reach an efficient classification process by their own, nonlinear combinations of these values are added as input of AdaBoost algorithm.

According to the classic voting approach, a final hypothesis resulting formula is implemented: The final vote is computed as the weighted sum of the classifiers selected by the AdaBoost training stage.

Instead of using two classes only (pedestrians and non-pedestrians) three classes are here used:

- Pedestrians
- Non-pedestrians
- Appearing pedestrians

Appearing pedestrians are pedestrians that are initially not completely visible, that is, partially occluded by obstacles, so only a part of the pedestrian's shape can be framed – the upper or side part only.

AdaBoost was trained using image windows determined by the previous steps of the algorithm. Candidate selection is a complex and critical step for the AdaBoost training process [16]. Both normal and flipped samples are used in the training process. If two or more images are very similar (i.e., the pixel-wise difference is less than a threshold) only one of them is used in the training process. Images framed in Italy, the Netherlands (during tests before the IEEE Intelligent Vehicle Symposium 2008 demonstration), and Korea were used in the training process.

Once a pedestrian partly hidden by a vehicle is detected – by the vision system only – the system issues an internal alert. In this case no warning is provided to the driver because the danger level is not yet determined.

When an alert is issued, the search continues in the same zone and tracking is started. As soon as a tracked pedestrian becomes visible also to the laser scanner, the direction of its movement is considered. If a pedestrian moving from the corridor's edge to the corridor's center is detected, a warning is then issued to the driver. Figure 16 shows an alert and its following warning.

It is of basic importance to note that the driver is warned only when the pedestrian is completely visible, like in other systems. However, the system presented in this work is more reactive than others, since the tracking starts when the pedestrian is only partially visible by one of the sensors. In an urban situation like this, promptness is an important key to the success of the system.



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 16**

Two subsequent frames of a sequence: **(a)** the system detects the partly occluded pedestrian and issues an internal *alert*, but no warning to the driver. **(b)** When a detection of a fully visible pedestrian follows the internal *alert*, a *warning* is promptly sent to the driver

Figure 17 shows some results obtained in complex conditions (bad weather condition, poor light condition) and a misclassification: a suddenly appearing pedestrian detected as a normal pedestrian.

### Automatic Braking

Previous works have introduced the Region of No Escape (RONE [44]) as the area that will be anyway occupied by the vehicle in the near future even if the driver steers away or brakes.

Like traditional pedestrian detection systems, pedestrians present in the RONE are here detected by laser and vision fusion; the presence of a pedestrian in the RONE triggers the automatic braking, as a Collision Mitigation system (i.e., when the accident is unavoidable).

The RONE is indeed important to detect an inevitable crash, but its size and shape are defined assuming an optimal driver maneuver. Most drivers, anyway, have (1) longer reaction times and (2) suboptimal skills; therefore an extension of the RONE is mandatory in order to act as a driving aid (avoid the crash) rather than a collision mitigation system.

It has to be understood, though, that any extension of the RONE will include areas in which the vehicle has a high probability to move, but it is not 100% sure to be occupied by the vehicle in the near future (e.g., in case of a very skilled driver implementing an optimal avoidance maneuver).

The system presented in this work defines the *deceleration area* as a suitable RONE extension designed for the specific scenario addressed by this project. In fact, pedestrians appearing from behind an obstacle might induce a reaction time longer than the time required by a driver to locate a clearly visible pedestrian. Therefore, still assuming that the driver is not distracted and has optimal driving skills, longer reaction times would enlarge the RONE.

The deceleration area width is defined to be equal to vehicle width and its length proportional to vehicle speed, but anyway bounded to 40 m, while yaw is considered constant for the trajectory computation.

Whenever a pedestrian appearing from behind an obstacle crosses the road and enters in the deceleration area, the system will activate the automatic braking, even if a skilled driver might be capable of avoiding it. This is why it is extremely important to be able to issue a warning to the driver well in advance, and provide sufficient time to perform a manual maneuver. The driver has indeed a more subtle and deeper perception of the pedestrian behavior and a manual countermeasure might be more effective than an automatic one.

In this section the concepts of driving corridor, RONE, and deceleration area are introduced. It is important to note the difference between these three areas in front of the vehicle, which are used for different purposes. As shown in Fig. 18, the driving corridor is a jagged area edged by static obstacles present in front



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 17**

Some suddenly appearing pedestrians correctly detected: (a) in an underground parking, (b) in the rain, (c) behind a misaligned vehicle, (d) behind a wall, (e) at night; and (f) a suddenly appearing pedestrian detected as non-dangerous pedestrian (false negative)

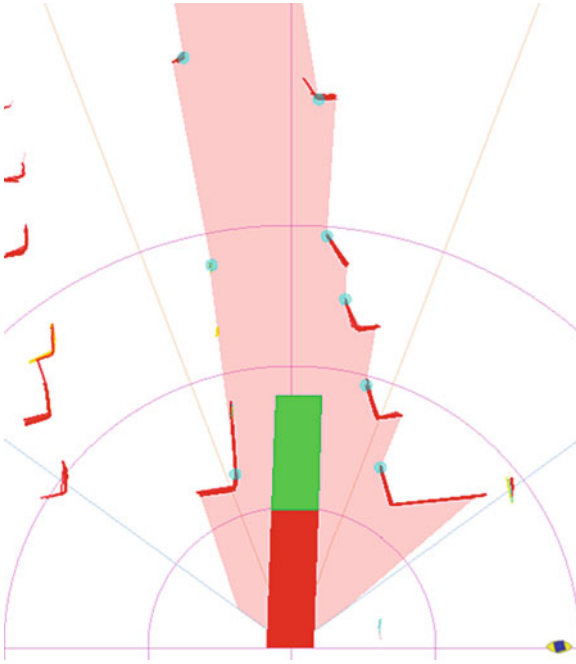
of the vehicle and represents the area where the vehicle can move.

The deceleration area is a smoother area that represents the positions where the vehicle should move considering a fixed yaw rate.

The RONE represents the area where the vehicle will move in the near future.

## Results

During the development of an ADAS it is important to consider different aspects for the results assessment: The first aspect is the precision of results, while the second one is the time that is needed to process the information; this second aspect is very important because the whole data processing must be operated real time.



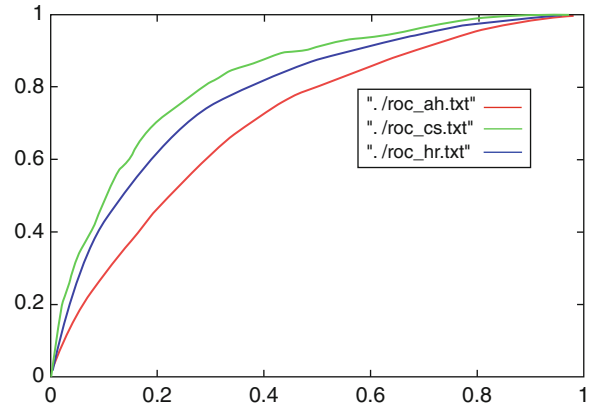
**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 18**

Driving corridor, RONE, and deceleration area: the *pink area* represents the driving corridor, the *red area* represents the RONE, and the *green area* represents the deceleration area

To evaluate the obtained results the parts of the system can be considered separately, in order to assess the performance of each part and to plan following developments.

The system can be divided into two parts, the first one is the regions of interest localization, while the second one is the search for pedestrians inside the region of interest.

The first part is simpler, and the results are only qualitatively evaluated. Regions of interest can be correctly detected with an appropriate precision up to a distance of about 40 m. Unfortunately some errors can occur due to strong vehicle pitch variation. If the laser orientation varies too much, the framed shape does not correspond to a significant representation of the environment. This problem is solved using a pitch rate sensor. If the pitch variation is too big, the whole system is disabled, because the first part of the algorithm is not able to get correct results. Anyway



**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 19**  
ROC curves for daytime classifier algorithm

generally high pitching is time limited, therefore the system is blinded for a short time interval.

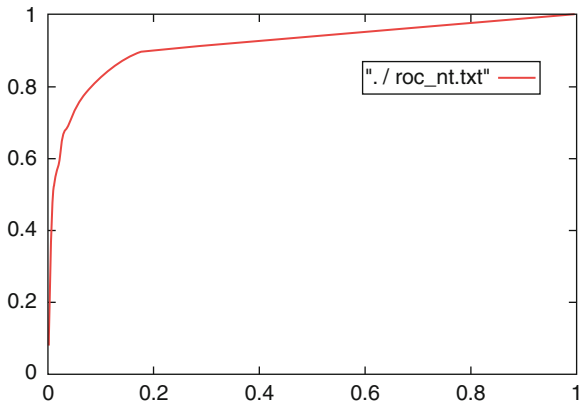
Search for pedestrians inside regions of interest is tested with three different test sets: one for daytime scenes, one for nighttime scenes with poorly illuminated environments, and one for nighttime scenes with proper illumination.

Figure 19 shows ROC curves for ad hoc features, Haar features, and cascade method. The algorithms based on Haar features reach more precise results than algorithms based on ad hoc features, but it is important to note that the cascade method, that merges the two methods, can reach a good classification rate. Results obtained on nighttime scenes with proper environmental lighting are similar to the previous one. Some tests also pointed out that the use of specific nighttime algorithms produce worse results.

The nighttime search, based on skeletonization, reaches good results without any further processing, as shown in Fig. 20.

Even if these results are important to develop the system, the final goal of the algorithm is not the detection of all pedestrians present in the scene, but the detection of dangerous pedestrians only. It is not important to detect a pedestrian normally walking on the sidewalk, while it is mandatory to detect a pedestrian that suddenly appears in front of the vehicle; therefore the results must be assessed considering the dangerous situations only.





**Active Pedestrian Protection System, Scenario-Driven Search Method for. Figure 20**  
ROC curves for daytime classifier algorithm

A set of sequences were considered for extensive performance assessment, for a total driving time of about 10 h in complex urban scenarios. A total of 236 km were driven during day and night, in different weather conditions (sunny, cloudy, rain, fog). Various scenarios were included: downtown, large and narrow roads, underground car parks, highways, and rural roads. During night tests, scenarios with and without external illumination were acquired. During night tests the NIR illuminators were always turned on. To complete the test sets some specific situations were staged (such as very dangerous pedestrian crossings), but a number of dangerous scenarios were anyway framed in normal driving.

In the tests, the following performance indexes were considered:

- Number of pedestrians suddenly appearing in front of the vehicle (that must generate warnings to the driver)
- Number of pedestrians appeared in front of the vehicle that have been successfully tracked and which can be hit by the vehicle (that must trigger the automatic brake)
- Number of fully visible pedestrians which can be hit by the vehicle (that must trigger the automatic brake)

For each performance index the number of correct detections, false positives, and false negatives were computed.

| Performance index | Correct detections | False negatives | False positives |
|-------------------|--------------------|-----------------|-----------------|
| A                 | 24                 | 11              | 1               |
| B                 | 5                  | 0               | 0               |
| C                 | 8                  | 0               | 0               |

Results of a 10-h drive (~236 km, ~540,000 images)

A total of 24 suddenly appearing pedestrians were correctly detected in the tests (case A). Only one false positive is present.

Only 1 missed warning out of 11 is due to a pedestrian miss-detection; the others are caused by either alert misses or delayed detections, but in these cases, even if the warning signal is not promptly issued to the driver, the automatic brake would have stopped the vehicle avoiding the crash if the pedestrian would have been in a dangerous situation. All pedestrians into the deceleration area are localized (no false negatives): five fall into case B, while eight fall into case C.

No false positives are present, thanks to the special attention paid to develop this SDS approach.

Anyway it is important to note that, for safety reasons – given that the tests were performed in real traffic – the size of the deceleration area was increased but, nevertheless, the system behaved very satisfactorily.

The limited number of dangerous events requiring the intervention of the safety system is not surprising. During normal driving dangerous situations are not frequent; plus, some of them are staged in order to challenge the system.

The analysis of the results obtained during the test highlighted good results also in case of rain, when pedestrian with umbrellas were detected as well. The system is anyway not able to discriminate between multiple pedestrians moving together or in situations in which the laser scanner is not able to obtain a clear picture, for example, when pedestrians hold bags or other large objects; moreover suddenly appearing pedestrians walking very slowly or pedestrians under very critical lighting conditions can be missed.

A particular attention was paid to the processing time. A system that it is not able to process the data real time is useless for automotive applications.

The great benefit of the SDS method is the reduction of the regions of interest that must be considered

by the vision system. Some tests focused that the number of considered areas is almost half of the ones considered using a traditional system. As the image processing is the most time-consuming part of the algorithm, the system can reach a significant reduction of global processing time.

Considering that the system is designed to work in urban and rural street, where the vehicle speed must be lower than 50 km/h, a processing frequency of 10 Hz can be considered sufficient.

Frame processing time changes according to the chosen method (nighttime method is quicker) and to the number of regions of interest which are analyzed. Anyway, the execution time on a good performing PC is always lower than 70 ms/frame (about 14 Hz), considering the whole system.

### Future Directions

An important concept in road safety aimed at avoiding collisions with vulnerable road users was presented. Instead of implementing a traditional pedestrian detection system followed by both a validation step and the assessment of the danger level, it was proposed an innovative approach. Whenever a specific environmental scenario is detected (i.e., stopped vehicles that may hide pedestrians), the possible presence of pedestrians is checked for in well-defined areas only, focusing on a particularly critical environment, typical of urban accidents.

Besides directly detecting dangerous situations, it also boosts timing performance, since the computationally intensive part – vision-based pedestrian recognition – is performed only on limited portions of the image.

The system was developed to tackle a very specific yet very common scenario, while it does not cover the whole set of dangerous situations that may occur in an urban environment.

The search for all the pedestrians present inside the so-called driving corridor can generalize the system, yet maintaining a scenario-driven approach, as the search is performed only in a corridor dynamically built using laser scanner data.

As a conclusion, the proposed approach offers the following advantages over traditional pedestrian detection systems.

- The system looks for specific spots in the 3D scene that might contain pedestrians (even partly occluded ones); in these areas the search for pedestrians is performed by vision; vision is therefore activated only in selected areas providing a lower number of false detections and a faster processing.
- Once a partly occluded pedestrian is detected, an internal alert is issued; when the pedestrian becomes completely visible, the initial alert is followed by a warning to the driver. Therefore the system reacts very quickly to the sudden appearance of a pedestrian, since tracking already started when the pedestrian was partly occluded and visible only to one of the sensors (vision).

This work was developed as a feasibility study aimed at determining whether the deployment of nonreversible systems might be a viable solution for real implementations. Nonreversible systems have always been considered extremely critical, since their robustness and effectiveness has to reach 100% to allow their integration on real vehicles. No sudden stops or irreversible maneuvers have to be triggered by false positives in the detection. Unfortunately, reaching a correct interpretation of all scenes and scenarios is as desirable as improbable and a lot of testing must be performed to evaluate possible failures in the detection.

### Bibliography

1. O'Farrell J (2006) May contain nuts. Black Swan, London
2. Lucchini E, Kramer M (1981) Safety device for the protection of pedestrians. Patent
3. Matsuura Y, Maki T (2001) Device for reducing the impact of pedestrians. Patent
4. Myrholt H, Ericsson M (2002) Deployable A-pillar covers for pedestrian protection. Patent
5. Buehler M, Iagnemma K, Singh S (eds) (2010) The DARPA urban challenge. Springer, Berlin
6. Agran PF, Winn DG, Anderson CL, Tran C, Del Valle CP (1996) The role of the physical and traffic environment in child pedestrian injuries. *Pediatrics* 1996(98):1096–1103
7. Retting RA, Ferguson SA, McCartt AT (2003) A review of evidence-based traffic engineering measures designed to reduce pedestrian-motor vehicle crashes. *J Public Health* 2003(93):1456–1463
8. Premebida C, Monteiro G, Nunes U, Peixoto P (2007) A lidar and vision-based approach for pedestrian and vehicle detection and tracking. In: *Proceedings of IEEE international conference on intelligent transportation systems*, Seattle, pp 1044–1049

9. Hwang JP, Cho SE, Ryu KJ, Park S, Kim E (2007) Multi-classifier based LIDAR and camera fusion. In: Proceedings of the IEEE international conference on intelligent transportation systems, Seattle, pp 467–472
10. Franke U, Heinrich S (2002) Fast obstacle detection for urban traffic situations. *IEEE Trans Intell Transport Syst* 3(3):173–181
11. Curio C, Edelbrunner J, Kalinke T, Tzomakas C, Von Seelen W (2000) Walking pedestrian recognition. *IEEE Trans Intell Transport Syst* 1(3):155–163
12. Gandhi T, Trivedi MM (2007) Pedestrian protection systems: issues, survey, and challenges. *IEEE Trans Intell Transport Syst* 8(3):413–430
13. Hodzic M (2008) Distronic plus and brake assist plus reduce rear-end collisions by 20%. <http://www.benzinsider.com/2008/06/distronic-plus-and-brake-assist-plus-reduce-rear-end-collisions-by-20/>. Accessed 30 May 2010
14. The Motor report (2010) Toyota Prius To ship with advanced safety gear. <http://www.themotorreport.com.au/29611/2010-toyota-prius-to-ship-with-advanced-safety-gear>. Accessed 30 May 2010
15. Juliussen E (2010) Driver assist systems. [http://www.autofocusasia.com/electrical\\_electronics/driver\\_assist\\_systems.htm](http://www.autofocusasia.com/electrical_electronics/driver_assist_systems.htm). Accessed 30 May 2010
16. [http://www.prevent-ip.org/en/prevent\\_subprojects/intersection\\_safety/intersafe/](http://www.prevent-ip.org/en/prevent_subprojects/intersection_safety/intersafe/)
17. BBC News (2010) Self-parking car hits the shops. <http://news.bbc.co.uk/2/hi/technology/3198619.stm>. Accessed 31 May 2010
18. [www.trafficcalming.org](http://www.trafficcalming.org)
19. Clarke A, Dornfeld M J (1994) Traffic calming, auto-restricted zones and other traffic management techniques: their effects on bicycling and pedestrians. national bicycling and walking study, Case Study No. 19, Federal Highway Administration 1994
20. Huang HF, Cynecki MJ (2001) The effects of traffic calming measures on pedestrian and motorist behavior. Research report FHWA-RD-00-104, Federal Highway Administration, U.S. Department of Transportation
21. Crandall JR, Bhalla KS, Madeley NJ (2002) Designing road vehicles for pedestrian protection. *Brit Med J* 324(7346):1145–1148
22. Yao J, Yang J, Otte D (2008) Investigation of head injuries by reconstructions of real-world vehicle-versus-adult-pedestrian accidents. *Saf Sci* 46(7):1103–1114
23. Koch W, Howard M (2003) Comprehensive approach to increased pedestrian safety in pedestrian – Car accidents. *Proc Inst Mech Eng Part D J Automobile Eng* 217:513–519
24. Fröming R, Schindler V, Kühn M (2006) Requirement engineering for active safety pedestrian protection systems based on accident research. In: *Advanced microsystems for automotive applications*, New York, pp 79–106
25. Schuster PJ (2006) Current trends in bumper design for pedestrian impact. SAE Paper No.: 2006-01-0464
26. Fredriksson R, Håland Y, Yang J (2010) Evaluation of a new pedestrian head injury protection system with a sensor in the bumper and lifting of the bonnet's rear part. <http://www.autoliv.com/wps/wcm/connect/5f997a004ce4f3f8ac68eef594aebdee/Pedestrian.pdf?MOD=AJPERES>. Accessed 30 May 2010
27. <http://www.ivss.se/>
28. Broge JL (2010) Autoliv actively saving lives. <http://www.sae.org/automag/techbriefs/11-2001/page2.htm>. Accessed 30 May 2010
29. Autoliv (2010) Pedestrian protection. <http://www.autoliv.com/wps/wcm/connect/autoliv/Home/What+We+Do/Recent%20Innovations/Pedestrian%20Protection>. Accessed 30 May 2010
30. Kerkeling C, Schäfer J (2005) Structural hood and hinge concepts for pedestrian protection. In: *Proceedings of 19th international technical conference on enhanced safety*, Washington, DC, pp 379–389
31. Gernóimo D, López AM, Sappa AD, Graf T (2010) Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans Pattern Anal Mach Intell* 32(7):1239–1258
32. Sun Z, Bebis G, Miller R (2006) On-road vehicle detection: a review. *IEEE Trans Pattern Anal Mach Intell* 28(5):1–18
33. Hussein M, Porikli F, Davis L (2009) A comprehensive evaluation framework and a comparative study of human detector. *IEEE Trans Intell Transport Syst* 10(3):417–427
34. Enzweiler M, Gavril DM (2009) Monocular pedestrian detection: survey and experiments. *IEEE Trans Pattern Anal Mach Intell* 31(2):2179–2195
35. Folster F, Rohling H, Meinecke M-M (2005) Pedestrian recognition based on automotive radar sensors. In: *5th European congress and exhibition on ITS and services*, Hanover
36. Ritter H, Rohling H (2007) Pedestrian detection based on automotive radar. In: *Proceedings of the international conference on radar systems*, Edinburgh
37. David K, Flach A (2010) CAR-2-X and pedestrian safety. *IEEE Veh Technol Mag* 3:70–76
38. <http://www.watchover-eu.org/>
39. <http://www.project-amulette.de/>
40. Broggi A, Cerri P, Ghidoni S, Grisleri P, Jung HG (2008) Localization and analysis of critical areas in urban scenarios. In: *Proceedings of IEEE intelligent vehicles symposium*, Eindhoven, pp 1074–1079
41. Leyrit L, Chateau C, Tournayre C, Lapresté J-T (2008) Association of Adaboost and kernel based machine learning methods for visual pedestrian recognition. In: *Proceedings of IEEE intelligent vehicles symposium*, Eindhoven, pp 67–72
42. Geismann P, Schneider G (2008) A two-staged approach to vision- based pedestrian recognition using Haar and HOG features. In: *Proceedings of IEEE intelligent vehicles symposium*, Eindhoven, pp 554–559
43. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the international conference on computer vision and pattern recognition*, vol 1. Hawaii, pp 511–518
44. Fuerstenberg K, Schulz R (2006) Laserscanners for driver assistance. In: *Proceedings of the international workshop on intelligent transportation*, Hamburg, pp 155–159

## Advanced Public Transport Systems, Simulation-Based Evaluation

HARIS N. KOUTSOPOULOS<sup>1</sup>, MOSHE BEN-AKIVA<sup>2</sup>

<sup>1</sup>Department of Transport Sciences, The Royal Institute of Technology, KTH, Stockholm, Sweden

<sup>2</sup>Department of Civil and Environmental Engineering, MIT, Cambridge, MA, USA

### Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Advanced Public Transportation Systems (APTS)

Simulation-Based Modeling of APTS

Microscopic Transit Simulation Example: MITSIMLab

Mesoscopic Transit Simulation Example: BusMEZZO

Future Directions

Bibliography

### Glossary

**APTS** *Advanced public transportation systems* is the transit component of Intelligent Transportation Systems (ITS). It refers to technologies for the collection, communication, and processing of data useful for the management of transit systems.

**AVL** *Automated vehicle location* systems combine vehicle location and communications technologies to automatically track the locations of a fleet of vehicles. AVL is an integral component of automated vehicle monitoring and control, fleet management, traffic signal priority, and other transit applications. AVL can be used to monitor schedule adherence, estimate arrival times, and communicate location data to a transit operations control center (TOC) or to field-installed devices that require real-time vehicle location data.

**APC** *Automatic passenger counters* are systems that count passengers as they board and alight the vehicle at a stop. APCs may be used with AVL systems in order to record the spatial distribution of passenger demand along a vehicle's route. APC technologies include treadle mats, infrared beams, and emerging computer imaging (still at early stage

of development). APC records the stop location, the time and date of arrival at the stop, the time the doors open and close, the number of passengers boarding and the number of passengers alighting.

**BRT** *Bus rapid transit* is an evolving transit concept aiming at combining the quality and capacity of rail transit and the flexibility of bus transit.

**Electronic fare payment** *Electronic fare payment* includes a range of technologies, such as smart cards, designed to reduce costs associated with fare collection and to improve customer convenience (also known as Automated Fare Collection (AFC) systems).

**Fleet management** *Fleet management* applications refer to "vehicle-based" technologies that may be used to improve vehicle planning, scheduling, and operations.

**ITS** *Intelligent transportation systems* is an umbrella term referring to sensor, communication, and computing technologies for improved management of transportation systems. ITS relates to all transportation modes.

**Simulation** "... the process of designing a model of a real system and conducting experiments with this model for the purpose of understanding the behavior of the system or evaluating various strategies for the operation of the system" [1].

**Signal priority** Transit *signal priority* involves the modification of a traffic signal's regular timing plan to give preference to transit vehicles. Signal priority is designed to reduce transit vehicle delays at signalized intersections.

**Traveler information** *Traveler information* refers to technologies designed to provide pre-trip and en route information to travelers to allow them to make informed trip-making decisions.

**Transportation demand management** *Transportation demand management* refers to systems aimed at improving the utilization of existing transportation network infrastructure, through influencing the demand characteristics.

### Definition of the Subject and Its Importance

According to the latest UN report on urbanization, more than 50% of the world's population lives in cities [2]. In Europe and the USA, it approaches 70%.

In part due to developments like these, congestion in urban areas continues to grow and create a number of negative impacts, including worsening air quality and noise pollution, consumption of scarce resources, lost productivity.

Public transportation is an important component of the transportation system and potentially a critical element of any strategy toward sustainable mobility in urban areas. The importance of improved public transportation services toward sustainable and efficient transportation systems is well recognized. With the emergence of Intelligent Transportation Systems (ITS), advanced sensor, communications, and computing technologies have been introduced to facilitate better planning, management, and control of public transportation systems [3]. Known as Advanced Public Transportation Systems (APTS), these technologies aim at developing public transportation systems that are more reliable and responsive to passenger needs, and hence, public transportation systems that offer higher level of service [4, 5].

APTS can be quite complex, with dynamic interactions among the various components. For that reason, simulation-based models that incorporate all the dynamic aspects, as well uncertainty inherent in the operations of public transportation systems, have drawn a lot of interest for evaluation and design of APTS.

## Introduction

APTS are ITS technologies applied to public transit in order to improve operational efficiency, cost savings, safety, quality of service, and other transit measures of performance. Some APTS applications offer potential for improving service by providing greater leverage to service providers for managing and controlling bus transit operations. Other APTS applications provide benefits in terms of speed, security, and convenience directly to the customer. APTS have the potential to significantly change the way transit services are provided and the way customers use them. Increasingly, popular technologies such as Automated Vehicle Location (AVL) systems, Automatic Passenger Counters (APC), and Electronic Fare Payment already have a wide range of impacts on bus transit operations.

The first examples of APTS date back to the late 1960s and early 1970s with the introduction of Automated Vehicle Monitoring (AVM) systems. The majority of these vehicle location technologies were signpost-based systems (with stationary signposts along bus routes). The signposts were equipped with electronic transmitters that emit unique identification codes. When a bus passed the signpost, an in-vehicle unit received the signpost's identification code and record the time and date, the difference between the current odometer reading and the last (recorded at the previous signpost), and the vehicle's identification code. The bus sent the information to the TOC via radio or other medium periodically or when prompted by the transit operations control center (TOC). Early implementations of APTS were expensive to install, operate, and maintain. Since then, new locations and communication technologies (e.g., GPS) have emerged and led to both improved performance and reduced costs.

## Advanced Public Transportation Systems (APTS)

Existing and emerging APTS applications of interest to this report can be grouped into [4, 5]:

- Fleet Management
- Traveler Information
- Electronic Fare Payment
- Transportation Demand Management

*Fleet Management* strategies focus on improving the planning, scheduling, and operations of a fleet of vehicles [6, 7]. The related strategies aim at improved service reliability, safety, and operating efficiency (e.g., reduced nonrevenue time, increased productivity) and faster service disruption recovery. In general, fleet management includes technologies that collect and make available vehicle performance data (e.g., vehicle location), and technologies that use that data for real-time control or for planning and scheduling. Important technologies used for fleet management purposes include: Communications Systems, Geographic Information Systems (GIS), Automated Vehicle Location Systems (AVL), Automatic Passenger Counters (APC), Transit Operations Software, and Traffic Signal Priority [8].



*Traveler Information* in transit applications refers to technologies used to provide travel information to passengers in order to assist their trip-making decision. The information provided may range from static route, schedule, and fare information to real-time vehicle location and/or estimated arrival time. AVL systems are the enabler for the provision of real-time information. Traveler information may be disseminated through various means, for example, web-based itinerary planning services, smart phones, in-vehicle units, etc. Traveler information is generally expected to improve the quality of transit service by improving the passenger experience. Traveler information may give passengers a better sense of control over their trip-making decisions and/or enable them to take action to minimize waiting times at stops, plan transfer connections, and thus reduce overall travel time. Information may be provided prior to departure (e.g., by phone, internet), at the terminal or stop, or in the transit vehicle. Information influences passenger trip-making decisions, for example, route, stop, departure time choice, etc. The attractiveness of transit alternatives is a function of waiting time, in-vehicle travel time, transfer time to the connecting trip, number of transfers, onboard comfort (crowding), etc. Information systems can provide timely information regarding many of these attributes [9, 10].

*Electronic Fare Payment* technologies forego cash and token payment with the aim of reducing the operating costs of fare collection systems, increasing safety and security on the vehicle, improving data collection, and increasing customer convenience [11, 12]. There are several electronic fare payment technologies, including magnetic stripe cards, and smart cards. Electronic fare payment technologies can have significant impacts on transit operations. The most obvious of the potential impacts on operations occur at the bus stop, where passengers board and alight from the vehicle. Depending on the type of electronic fare payment technology, considerable gains can be realized in terms of lower dwell times at stops through increased transaction speeds. Contact card technologies, where the card is physically swiped through a card reader, and contactless card technologies, where the card and card reader communicate without physical contact but rather via an electromagnetic signal, affect passenger

boarding rates differently. Boarding rates increase to a greater extent with contactless card technologies.

*Transportation Demand Management* is the application of technology to alter the usage patterns of the transportation network, with an emphasis on encouraging users to travel by transit. There is a broad range of technologies designed to better coordinate various transit services, and provide forums for organized carpooling and carsharing. In general, approaches to managing transportation demand aim at increasing the number of high occupancy vehicles in congested transportation networks, promote travel in off-peak hours, and implementing transit incentive programs. Examples include dynamic ridesharing and automated service coordination [13].

## Simulation-Based Modeling of APTS

### Overview

From the previous discussion, it is clear that evolving information, communications, and sensor technologies and innovative transit operations control strategies are becoming critical elements of viable, competitive public transit systems. APTS, through a variety of data collection and communication capabilities, support improved operations planning and real-time transit operations management. APTS, for example, provide valuable input to planning applications that may lead to better transit system design (e.g., improved scheduling and route planning). APTS also enable a variety of real-time operational strategies (e.g., holding and dispatching) that aim at schedule maintenance and schedule restoration (after major disturbances). Additionally, APTS allow the sharing of real-time performance information with travelers to influence demand and improve passenger level of service.

As innovative technological solutions are integrated with transit services and becoming more common, it is important to have appropriate tools for evaluating and refining the corresponding operating strategies prior to deployment. Because of the complexity of the system, both from the supply and demand sides, simulation is an appropriate tool to evaluate, refine, and design APTS strategies. Simulation allows the representation of the dynamic aspects of transit operations as well as capturing of the stochastic behavior of the system. Hence, simulation enables representation of complex



interactions between the network components: general traffic, transit vehicles and passengers, and comparison of various scenarios and alternatives under realistic and consistent conditions.

In the context of general traffic operations, simulation models have been developed and used for evaluation and design since early on [14]. Traffic simulation models, based on their level of detail, are characterized as macroscopic, mesoscopic, or microscopic. Macroscopic models use a fluid approximation to capture traffic dynamics with a continuous representation of traffic flows [15, 16]. Mesoscopic models represent individual vehicles (which have travel behavior associated with them, e.g., route choice) [17–19]. However, the modeling of traffic dynamics is rather aggregate through speed-density relationships and queuing theory considerations. Microscopic models represent individual vehicles which in addition to travel behavior also have driving behavior. The network is represented in detail, and vehicles make acceleration/deceleration and lane-changing decisions based on traffic conditions, information, and overall travel plans [20, 21]. Hybrid simulation models have also been developed that represent different parts of the network and the associated operations at different levels of details. They combine microscopic simulation at small areas of interest, while the remaining network is represented at a mesoscopic level of detail [22].

Transit simulation has not been developed at the same pace as pure traffic simulation models. Although simulation models can have many advantages for public transportation research and applications, comprehensive transit simulation models are still at an early stage of development. An extensive review of 32 microscopic traffic simulation models report that only 52% represent public transportation at all, 42% model transit priority, and only 6% model transit traveler information systems. At the same time, the majority of users they interviewed were interested in large-scale applications at the urban or regional context, and that these users ranked modeling of public transportation the second most important capability in traffic simulation models [23]. Another review of a representative group of microscopic simulation models concluded that a detailed representation of bus operations and the interactions between buses and other vehicles has been largely ignored or unrealistically simplified [24].

In the context of microscopic simulation models, most of the early advances related to transit systems have focused on implementation of transit signal priority [25–28] or operations of bus stops [24, 26, 29]. For example, in [26], the microscopic simulation model DRACULA was enhanced, mostly from the traffic point of view. The transit modeling was limited, and it did not include the representation of schedules (arrival and departure times at each stop). Passenger arrival rate was drawn from a normal distribution with a stop-specific average value and fixed variance. The dwell time was a function of the number of boarding passengers only. Targeted applications included the evaluation of roadside vs. bay stops, reserved bus lanes vs. guide ways (special ways for guided buses), and bus signal priority.

Recently, several microscopic traffic simulation models have significantly enhanced or extended their transit capabilities, indicating a shift away from traditional, simplistic bus transit modeling and an increased emphasis on realistic bus transit operations simulation. The microscopic traffic simulation model MITSIMLab was systematically extended to represent in detail APTS operations [30]. Earlier work with DRACULA was extended by integrating a transportation planning tool [31, 32]. The simulation includes supply and demand submodels that interact according to a learning algorithm, where the demand (in terms of drivers' route choice) for each day is affected by the cost experienced in the preceding day. The simulation is based on two loops: external (day-to-day) iterations of the demand loop and internal (within-day) iterations of the supply loop.

Similarly, commercial packages such as Transmodeler [33] and VISSIM [34] have added extensive transit simulation capabilities. A number of more specialized systems have also been built on top of available commercial traffic simulation tools. For example, as Bus Rapid Transit (BRT) systems gain in popularity, simulation models of their operations have also appeared in the literature [35]. SmartBRT, a microscopic simulation model for evaluating BRT performance, is built on top of PARAMICS [36]. PARAMICS is a microscopic traffic simulation model [37]. The level of transit integration is rather low since the SmartBRT entities, including bus stops and passengers, do not interact directly with corresponding

PARAMICS entities. This can create considerable complications and inaccuracies in transit modeling and is indicative of the difficulties in integrating successfully transit capabilities on top of existing tools. The model has the capability to represent bus signal priority, various fare collection mechanisms, and detailed dwell-time functions.

MISTRANSIT (Microscopic Simulation Transit) is a PARAMICS API for modeling transit operations [38, 39]. In general, the movement of the buses is run by PARAMICS, while MISTRANSIT operates control strategies and collects transit-related measures of performance. Detailed representation of dwell times at stops as a function of vehicle design parameters, and boarding and alighting passengers are included. Demand is represented through origin–destination flows between stops resulting to individual passengers with attributes, such as stop arrival time, and transfer parameters. MISTRANSIT was used to evaluate holding strategies, operations at bus stops nearby traffic signals, bus signal priority, and capacity of busways with skip–stop operations and interchanges between various public transportation components.

MILATRAS is another API that adds dynamic transit assignment capabilities to PARAMICS [40–42]. It aims at analyzing transit system performance and level of service under real-time information, as well as evaluation of BRT and LRT systems. Given that the underlying platform is a microscopic simulation model, MILATRAS is suited for the analysis of medium-size networks with medium to low frequencies. The model adds dynamic representation of individual passengers and their learning process as they adjust their behavior in response to changes in transit system's performance. The transit path choice problem is modeled as a Markov decision process with each path decision (departure time, boarding and alighting decisions) expressed as a state-action pair. While the overall model does not seek an equilibrium solution, the Markov decision process fulfills the conditions for convergence to a unique and optimal solution (in terms of passenger state-action choices).

Microscopic models are limited when applied to large-scale applications because of the level of detail and extensive computational effort they require. In contrast, mesoscopic simulation models, which represent individual vehicles but avoid detailed modeling of

their second-by-second movement, are more appropriate for system-wide evaluation of transit operations and APTS, as they are for general traffic. Research in this area is also limited, especially with respect to APTS applications with large-scale networks [43].

However, activity has recently increased with research focusing on both representation of APTS in mesoscopic simulation models and development of demand models that capture transit-related travel behavior. In contrast to microscopic models, mesoscopic models have a greater need for detailed representation of passenger demand. In this respect, dynamic transit assignment capabilities are an important feature of mesoscopic transit models. In order to evaluate the long-term effects of bus signal priority strategies on drivers' route choice and modal shift, a simulation-based dynamic traffic assignment model within DYNASMART was developed [44]. As a mesoscopic traffic model, DYNASMART represents individual vehicles and uses speed-density relationships to capture traffic dynamics [18]. Transit vehicles are generated deterministically, and dwell times are not calculated explicitly but taken as a fixed delay on the link capacity. Passengers are generated according to time-dependent origin–destination zone demands. Passengers have individual preferences and characteristics that impact their behavior.

A comprehensive transit modeling framework was proposed in [45]. The framework uses a multi-agent approach to transit simulation, focusing on improving the representation of travelers' behavior. The public transportation system consists of three components: passenger behavior, road traffic dynamics, and specific bus-network operations, which include the interactions between the buses, passengers, and road traffic. The network representation is multimodal including pedestrian, road, and bus networks. The multi-agent approach considers the roles that each agent plays: a bus plays two roles simultaneously (vehicle and transportation service), while a traveler plays two roles alternately (as a pedestrian and as a bus passenger). While the vehicles' schedules and travelers' routes are predetermined, their progression is determined by the interaction between the three components of the system. The simulation implementation uses a hybrid approach, where buses and travelers are simulated at a microscopic level of detail and

remaining vehicles are simulated at a macroscopic level of detail.

Another comprehensive framework for mesoscopic transit simulation is BUSMezzo [46, 47], capable of modeling transit operations, and evaluating and designing APTS at the network level. The transit simulation model is fully integrated within the platform of Mezzo, a mesoscopic traffic simulation model [48]. Mezzo is an event-based simulator and is designed with hybrid capabilities (to interface with microscopic models).

### General Requirements for Effective Transit Simulation

In order to simulate APTS applications, it is necessary to represent bus transit operations at a level of detail that supports the operational characteristics of the technology or system of interest. For example, AVL systems that monitor bus performance and determine holding and dispatching solutions to schedule deviation problems cannot be simulated in a model that does not represent explicitly bus schedules. Figure 1 summarizes various APTS technologies and the implications they have on bus transit operations and their representation in a simulation framework. In general, APTS

- Enable a variety of real-time operational strategies (e.g., holding and dispatching) that directly affect transit vehicle movements
- Provide valuable input to planning applications that lead to better transit system design (e.g., improved scheduling and route planning)
- Allow the sharing of real-time performance information with travelers to influence demand and improve passenger level of service (e.g., route choice)

In order to simulate APTS applications, the simulation model should be based on a detailed representation of bus transit operations. Bus operations are subject to a number of incidental and controlling forces, including service schedule design, passenger demand, and dispatcher control and intervention. A simulator capable of capturing the behavior of these interacting components should have the following elements:

1. *Transit Supply* representation at the appropriate level of detail. Transit supply refers broadly to the

system level components of bus transit operations that are generally under the control of the transit service provider. The transit supply includes the transit network, schedules, and fleet assignments. It also includes the vehicle movements and interactions with the remaining traffic, as well as the representation of APTS. APTS representation is crucial and includes surveillance and monitoring systems that generate and distribute real-time information, the application of that data to real-time control strategies, etc.

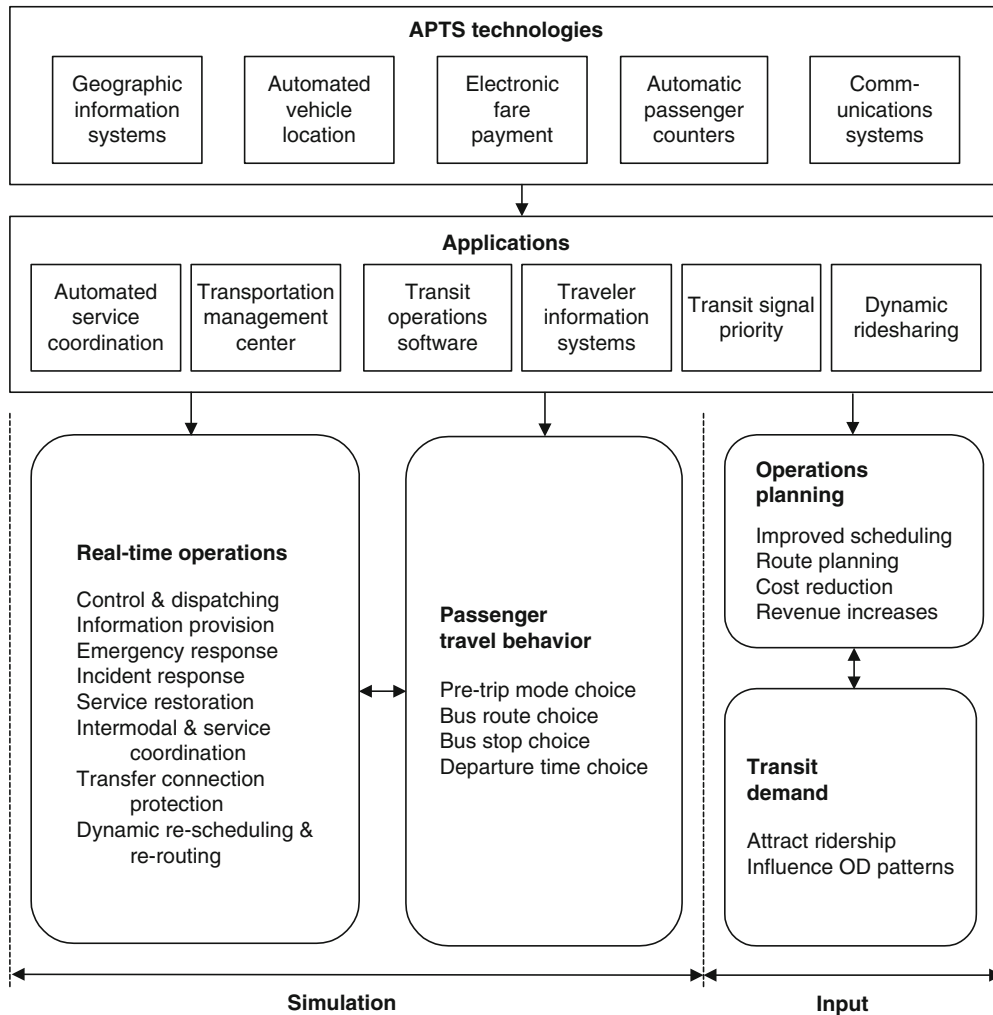
2. *Demand representation*. Transit demand refers to the passengers and their behavior with regard to use of the system, including en route and pre-trip decisions, such as route choice, as well as behavior at bus stops and transfer points.
3. *Measures of effectiveness (MOE)*. MOEs include the indicators, levels of service, and other measures that are used to evaluate the performance of a transit system or an APTS strategy.

The model requirements may be incorporated into a simulation model by providing system variables and parameters as input to the model and/or modeling internally the effects of APTS on system variables. Some transit system variables, such as routes, schedules and passenger demand, are strongly influenced by operations planning or passenger trip planning applications and do not generally vary during the course of a simulation period. These variables are typically represented as static inputs to the model. APTS that interact with real-time operations, on the other hand, should be represented using models that capture interactions between:

- Buses and other modes
- Buses and passengers
- Buses and field-installed control devices
- Bus operators and the operations control center
- Passengers and traveler information systems

### Microscopic Transit Simulation Example: MITSIMLab

Microscopic simulation models are very detailed and appropriate for evaluation of APTS strategies at the operational level, including the detailed representation of various design aspects and parameters. MITSIMLab



**Advanced Public Transport Systems, Simulation-Based Evaluation. Figure 1**  
 APTS impacts on transit operations and implications for simulation (source: [30])

is an example of a general microscopic traffic simulation laboratory that models APTS in detail [30].

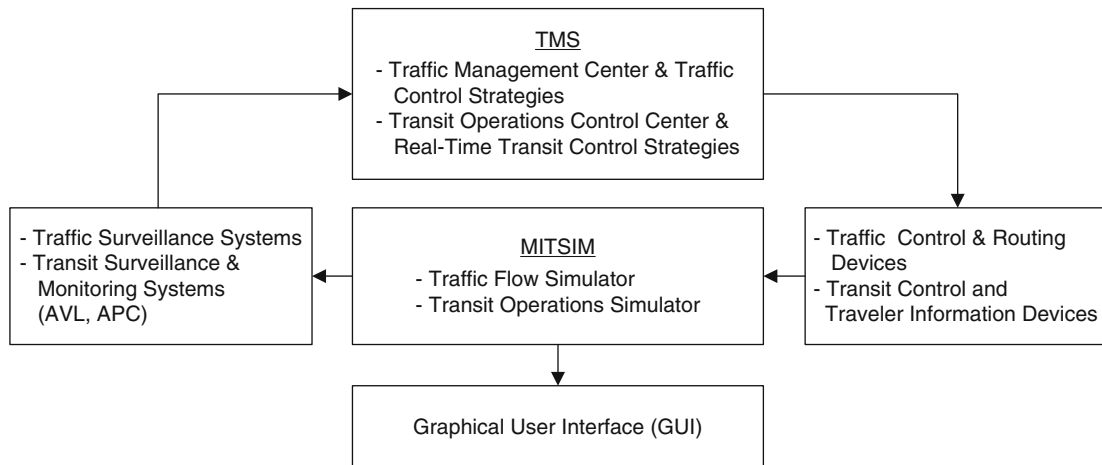
### Transit System Simulation

MITSIMLab was originally developed for the design and evaluation of advanced traffic management systems (ATMS) and advanced traveler information systems (ATIS) [49, 50]. It is based on sophisticated behavioral models, including driving behavior and travel behavior (e.g., route choice), that capture the complex decisions that drivers make before departing and en route to their destination. MITSIMLab has

a modular structure, which facilitates the addition of new functionality to the model.

The objectives of the overall bus transit model are to simulate, at a high level of accuracy and detail, the continuous and complex interactions between vehicles of various types, surveillance technologies, traffic management strategies and control, and information dissemination devices in the network. The modular framework of MITSIMLab, enhanced with bus transit simulation, is illustrated in Fig. 2 with its two main components: MITSIM and TMS.

MITSIM represents the physical road network and simulates the movements of individual vehicles of all



**Advanced Public Transport Systems, Simulation-Based Evaluation. Figure 2**  
Bus transit and APTS modeling framework (source: [30])

types through the network. As such, MITSIM represents the physical portions of the transit network (e.g., bus stops and bus lanes) and simulates the movements of transit vehicles and the interactions between transit and non-transit vehicles. MITSIM requires detailed input describing the transit and traffic network, bus operations, and travel demand (transit and car).

**Transit Supply Representation** The components of the transit system (transit network, schedule design, and fleet assignment) are generally considered to be static information, provided as input. A detailed representation of routes and schedules is used such that transit and traffic operations in the simulated network are sensitive to the variation in the route and schedule inputs. The transit system representation is based on the following inputs: network, vehicle movements, transit surveillance and monitoring, and transit operations control center.

**Transit Network.** The transit network representation describes the elements of a transit network, including bus stops and road segments that are used by transit routes. Routes are defined as sequence of links that make up the path and a list of bus stops along the path. Additionally, the transit network representation supports the simulation of deadheading trips, trips between depots and terminals, and other vehicle movements.

- **Schedule Definition.** The schedule definition input describes trips on a route and their timing. When the schedule definition input is read, trips are created. Trips have a single assigned route and a series of scheduled arrival times at the stops or time points along that route. The representation of the schedule is flexible to allow the simulation of different types of routes (e.g., high frequency, express, etc.).
- **Run Definition.** The run definition input specifies the series of trips to which a bus may be assigned. When the run definition file is read, runs are created as a sequence of trips to which the bus is assigned. The representation of runs is generic to allow the simulation of interline, deadheading, and other nonrevenue trips.
- **Fleet Assignment.** The bus assignment input defines the run to which each bus in the fleet is assigned, the type of bus (e.g., articulated), and the departure rate, or time, at which buses enter the network to begin their assignments.

**Transit vehicle movements.** A set of models capture bus operator behavior which may be different than other drivers in the network. For example, bus operator travel behavior is a function of the route and schedule input rather than merely an origin and a destination. Similarly, bus operator driving behavior includes maneuvers and decision-making processes that are unique to bus operators. Bus operators may accelerate,

decelerate, change lanes, and perform other maneuvers in order to arrive at a bus stop, reenter the traffic stream from a bus stop, and achieve a variety of other objectives that do not pertain to other drivers. Furthermore, other drivers may change their behavior in the presence of buses. Some of the more important elements of vehicle movement and interactions that are modeled include:

- *Behavior between stops.* Operator behavior between stops refers to driving behavior that controls the vehicle's trajectory from one stop to the next. Some of the tasks of bus operators between stops include mandatory lane changing to arrive at bus stops, discretionary lane changing when the bus is far from the downstream stop, and lane choice when bus lanes are present. Furthermore, some bus operator characteristics capture the fact that bus operators generally know their routes very well and are familiar with the network.
- *Behavior approaching stops.* This pertains to both car drivers and bus drivers. The tendency of private auto drivers to overtake buses when traveling behind a bus in a lane that contains stops downstream as an important element of general driving behavior near stops was reported in [24]. It is postulated that drivers anticipate the bus' eventual deceleration to a stop to serve passengers and thus will change lanes or "squeeze" around the bus in the same lane in order to pass. This suggests that general car-following acceleration models do not capture the interactions between buses and other modes correctly. Accordingly, in MITSIMLab, drivers have a high probability of changing lanes when traveling behind a bus in a lane that contains bus stops downstream. That probability increases when the bus is actively decelerating to pull into a stop.
- *Behavior at Stops.* Behavior at stops deals mostly with dwell time, the time the bus operator spends stopped with the doors closed, the time the operator takes to open and close the doors, and the time it takes to serve passengers while the doors are open. The most common dwell-time models assume dwell time to be a linear function of the number of boarding and alighting passengers. A wide variety of dwell-time models have been proposed in the

literature, including models that account for fare collection type, bus type (e.g., low-floor, number of doors), and other factors [51–53]. The default dwell-time model in MITSIMLab is the one suggested in [53], the Transit Capacity and Quality of Service Manual (TCQSM).

$$T = \gamma + \alpha A + \beta B + \varepsilon,$$

where  $A$  is the number of alighting passengers,  $B$  is the number of boarding passengers, and  $\varepsilon$  is a random term. The values of parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  depend on bus type and design, fare payment strategy, and passenger characteristics, etc. A crowding factor may also be added when there are standing passengers on the bus [52].

- *Behavior departing from stops.* The critical bus operator maneuver near wayside bus stops is the departure from the stop. When a steady traffic stream occupies the general traffic lane into which the bus must enter, the operator must locate a gap into which it may accelerate from a standstill. General lane-changing models might not capture this behavior realistically. In most lane-changing models, the acceptable gap into which the subject vehicle will merge is a function of its position relative to the lead and lag vehicles and its speed relative to the lag vehicle in the target lane. For a bus departing from a stop, where its speed is zero, such a lane-changing model may not be adequate and could lead to unrealistic delays. Specially adjusted gap acceptance models are used to capture this behavior.

*Traffic surveillance and monitoring.* The systems in place to collect information on the status of the transit operations are simulated in detail (to the level of functional characteristics, e.g., measurement errors). Surveillance is a generic term that refers to various detection and sensing technologies located in the transportation network or onboard vehicles. Some APTS surveillance technologies include automatic vehicle location (AVL) and automatic passenger counters (APC). The communication between MITSIM and TMS is used to mimic the transmission of vehicle location and passenger load information between vehicles (MITSIM) and the transit operations control center or field-installed devices (Fig. 2).



There are three issues related to real-time bus transit surveillance data that are taken into consideration: ownership, storage, and distribution. *Ownership* refers to the entity (e.g., bus) that has possession of the data. For example, to emulate APC technologies, buses store passenger load information as passengers board and alight, and the user may specify input parameters to capture the random error in APC counts. *Storage* refers to the length of time an entity keeps the data before purging it, deleting it, or sending it to another entity. For instance, buses keep track only of their most current schedule information, which is updated each time they arrive at a stop. When a bus arrives at a sensor upstream of a signal equipped for priority, the sensor collects the schedule adherence information and relays it to the signal controller in TMS to determine whether priority should be granted. Finally, data *distribution* deals with when and between which entities the information is shared. If bus load and location information is collected periodically, as through AVL systems, then this information is reported to TMS according to the system logic. Through an accurate representation of data ownership, storage, and distribution, MITSIMLab is able to represent any underlying AVL or other bus surveillance system.

*Transit operations control (TOC) center activities.* The operations control strategies as well as decentralized field-deployed strategies are simulated in TMS. TMS mimics the logic of the strategy under evaluation, and may use real-time traffic and transit data from the surveillance system as input to that logic. As a TOC simulator, TMS handles incoming transit surveillance data and executes APTS operations that take place in a TOC.

Device-based control strategies, such as signal priority, are also simulated. The information regarding approaching buses received from the surveillance system is used to trigger the corresponding strategy which calculates appropriate actions and sends the signal timing data to the signal controller in TMS. TOC-based strategies, on the other hand, can be simulated by direct control of vehicle movements. For instance, stop-based control strategies, like holding of a bus if it is ahead of schedule, are simulated by placing conditions on bus departures from a stop. APTS control strategies can vary widely in design and implementation from agency to agency and from application to

application. The modular structure of the simulation model facilitates the testing of a large number of control strategies.

**Demand Representation** The representation of the passenger demand is an important element of a simulation model. Passenger demand in MITSIMLab is modeled at two levels of detail (in increasing order of data requirements):

- Dwell times are randomly generated independently for each stop according to probabilistic distributions. However, such a simplified representation of demand, although efficient in terms of input requirements, ignores the impact that dwell times at one stop may have on the headway and, hence, dwell times at downstream stops.
- Arrival rates and alighting rates (e.g., as a percentage of the bus load) at stops. The model assumes that passengers arrive according to some probabilistic distribution and generates randomly the number of passengers waiting to board the incoming bus based on the actual headway from the previous bus. Likewise, the percentage of the bus load alighting at a stop determines the number of alighting passengers. In this representation, passengers are generated in aggregate and have no identifying characteristics, origins, or destinations.

For the last representation of demand, it is important to model the way passengers arrive at bus stops. Three categories of transit passengers have been reported in the literature: those who arrive coincidentally with the bus and thus have no waiting time, those who are familiar with the schedule and arrive close to the vehicle arrival time, and those who arrive randomly [54]. The fraction of the passengers in each category depends on the frequency of service. For example, for high-frequency routes, it is reasonable to assume that passengers arrive randomly. The model uses an aggregate representation of demand with time-dependent arrival rates. The default arrival model assumes that passenger arrivals have a Poisson distribution:

$$P[B(t, t + T) = k] = \frac{(\lambda(t)T)^k e^{-\lambda(t)T}}{k!}$$

where  $P[B(t, t + T) = k]$  is the probability that  $k$  passengers arrive in the interval of length  $T$  and  $\gamma(T)$

is the average arrival rate during the corresponding interval.

**Measures of Effectiveness** An important requirement for any transit simulation model is the generation of appropriate measures of effectiveness (MOEs). In general, simulation allows the collection of a considerable amount of output data, including vehicle trajectories, sensor readings, and segment and link statistics. MITSIMLab collects detailed information (e.g., passenger load, speed, mileage) about each bus and at stops (e.g., dwell times, passengers boarding, alighting and left behind) at user-specified intervals during a simulation. From the detailed data, a number of MOEs can be generated to characterize the performance of the system as dictated by the nature of the application. These MOEs are organized around five levels: system, route segment, stop, vehicle, and passenger.

### Case Study: Evaluation of Transit Signal Priority

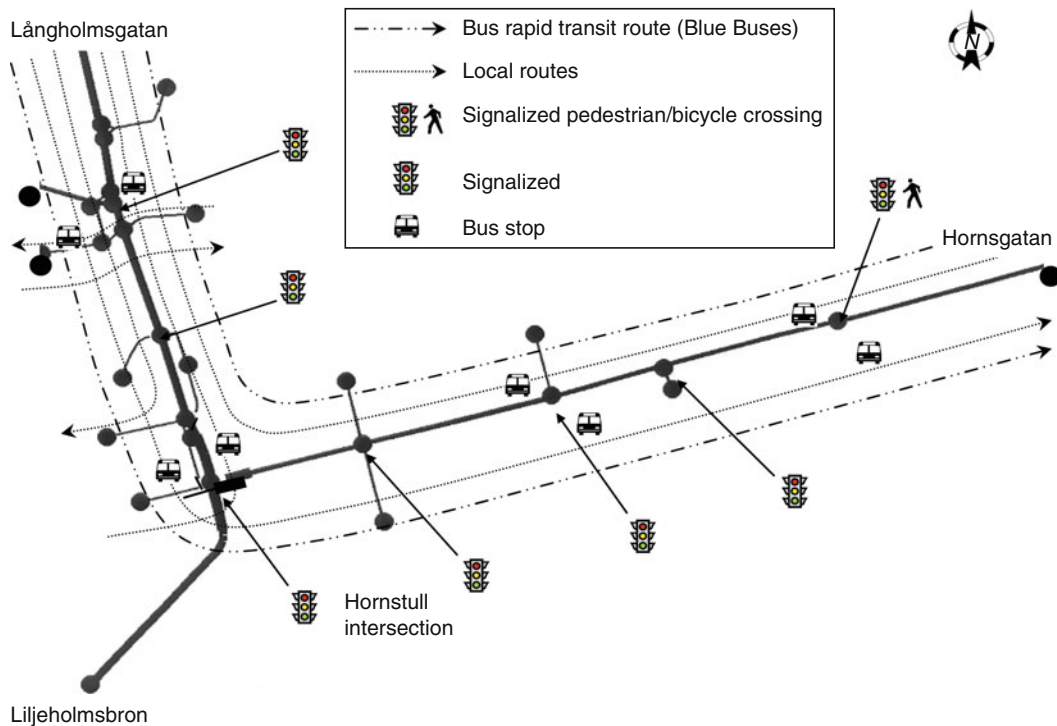
The purpose of this study is to evaluate the effectiveness of transit signal priority using the MITSIMLab laboratory (more information can be found in [30] and [55]). ITS, with the associated advances in sensor technologies, has enabled the development of interesting transit signal priority strategies [56–59]. Traffic signal priority is designed to reduce transit vehicle delays at signalized intersections. Reduced delay at intersections may also reduce overall travel time, aid schedule adherence and headway maintenance, and increase person throughput at the intersection.

Transit signal priority strategies are varied: they can be passive or active, and unconditional or conditional. Passive priority requires no communication between vehicle and controller, and involves the development of a signal timing plan that reduces delay on the transit vehicle's approach. Passive priority can be achieved by allotting more green time to the transit vehicle's approach, coordinating signals to improve progression along a bus corridor. Active signal priority dynamically adjusts the signal timing when the transit vehicle is detected. This is accomplished by extending the green interval in the current phase, terminating the current phase to start an early green interval for the transit vehicle's approach, or by inserting an extra green

phase on the vehicle's approach. As such, active priority requires sensors that identify approaching buses. Depending on the implementation, a communications system may also be needed to convey the sensor information to the traffic signal controller. The signal controller uses the information to adjust the timing of the signal plan to favor the movement of the approaching bus.

Active traffic signal priority can be unconditional or conditional. Unconditional strategies give priority to every qualified vehicle that approaches the intersection regardless of the vehicle's schedule or the impacts on the conflicting approaches. Conditional priority grants priority to approaching transit vehicles only if the vehicle meets some predetermined condition(s). In general, the condition for priority depends on the vehicle's current performance. Based on the available information, it can take various forms. For example, in the simplest case, the condition is based on the schedule (i.e., whether the vehicle is ahead of or behind schedule according to the timetable). However, as APTS facilitates the monitoring of the system in greater detail, other, more advanced, conditions can also be envisioned (based for example on the passenger load on the vehicle, headway from preceding bus, or other measurements or combination of criteria). Thus, communications and controller technologies that support conditional signal priority must be able to transmit and manipulate different pieces of data for evaluating priority eligibility.

In this case study, a number of conditional and unconditional signal priority strategies in an urban network in Stockholm, Sweden, are considered. The study network is located at the western end of Södermalm, one of a cluster of islands in inner Stockholm. The simulated time period of interest is the 7:30–8:30 a.m. morning peak hour. The study network is dominated by the Hornstull intersection that joins three major arterials: Liljeholmsbron, Långholmsgatan, and Hornsgatan (Fig. 3). There are six signalized intersections and one signalized pedestrian and bicycle crossing in the network. Three local bus routes and one BRT route operate within the study area. Buses that serve the local bus routes have 15-min headways during the peak periods, and the BRT buses operate at 7.5-min headways. BRT and local buses share all bus stops. The cross-street traffic in the study network is relatively low compared to traffic on the three arterials



Advanced Public Transport Systems, Simulation-Based Evaluation. Figure 3

Bus routes in the study network (source: [30])

due mostly to street closures that prevent drivers from using alternate routes to avoid congestion.

Only BRT buses are equipped to communicate with sensors to request priority. The objectives of the study are to evaluate the effects of various transit signal priority strategies on transit service and network performance. Four priority implementations are compared in the case study:

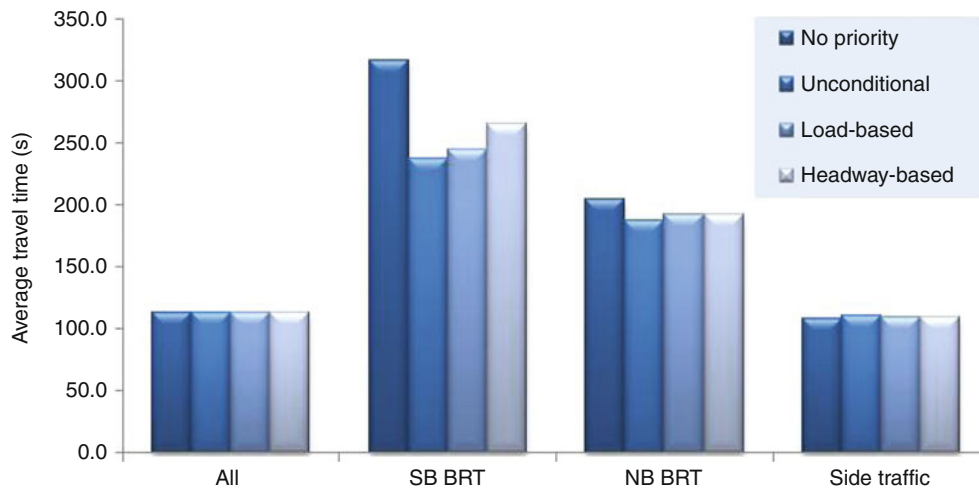
- No priority (base case)
- Unconditional priority
- Headway-based conditional priority
- Load-based conditional priority

For headway-based conditional priority, the time headway between the bus requesting priority and the bus before it must be equal to or greater than a minimum headway threshold in order for priority to be granted. A load-based conditional priority strategy gives priority to buses only when the number of the passengers onboard (load) exceeds a threshold. With the use of automated passenger counters (APC), such a strategy is feasible under APTS.

The analysis considers two benefits from each strategy: improved person throughput and reduced transit travel times. However, while it is of interest to assess the extent to which different signal priority strategies achieve these objectives, it is also important to evaluate other impacts, especially to the vehicles entering the network from the side streets. Hence, the measures of effectiveness include average vehicle travel time and total person travel time for all vehicles, BRT buses alone, and side-street vehicles alone. Impact on travel-time variability is also considered.

A sensitivity analysis designed to explore the effects of increased side-street demand is also conducted. Furthermore, given that APC technologies may suffer from measurement errors, the impact of such errors on the performance of signal priority was examined.

Vehicle travel time, total person travel time, and travel-time variability results for the priority strategies as well as the base case were collected using the outcome from several independent simulation replications. A priori, it is expected that BRT vehicles should experience greater travel-time savings, while non-priority



**Advanced Public Transport Systems, Simulation-Based Evaluation. Figure 4**  
Signal priority impact on travel times (source: [55])

movements (i.e., side-street traffic) should experience higher travel times as the conditions for signal priority become less restrictive.

Figure 4 summarizes the main results. Average BRT travel times decrease as the priority conditions become less restrictive, with the lowest average travel times occurring under unconditional priority. Since the demand on the side streets in the study network is very low, the impact on non-transit, non-priority vehicles entering the network from the side origins is not substantial.

Interestingly, the results also indicate that conditional priority strategies can achieve similar benefits as the unconditional priority strategy without granting priority quite as often. Under conditional priority about 70% of the buses receive priority. This tends to indicate that conditional priority strategies also have less impacts on the side traffic. However, due to the low demand on the side streets in this case, the change in average travel time for side-street vehicles for different priority strategies is too low to draw general conclusions about the trade-off between transit travel-time savings and side-street travel-time penalties.

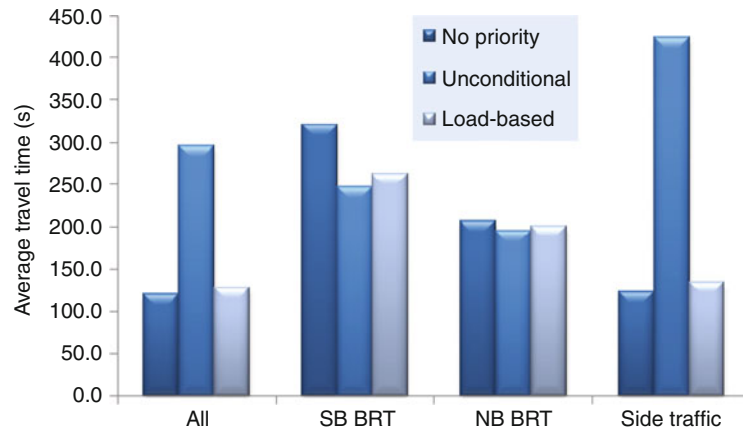
Total person travel time is the total time spent by all commuters in the system. The results show that under conditional priority, particularly load-based conditional priority, the difference between conditional and unconditional priorities is very small (compared to the difference with respect to average travel time per

vehicle). For example, the total passenger travel time for the load-based conditional priority is only 1.5% higher than the unconditional priority case.

Travel-time variability is an important measure of performance from the standpoint of traveler level of service. The results from the simulation study indicate that travel-time variability for BRT vehicles was also reduced under signal priority. Furthermore, conditional priority yielded almost similar benefits in that regard, as unconditional priority.

A major advantage of simulation is that it allows testing of a large number of different scenarios and facilitates sensitivity analysis with respect various operating and design parameters. In this case study, the transit simulation laboratory is used to evaluate the impact of increased demand levels as well as the impact of the accuracy of the APC equipment used to count boarding and alighting passengers.

Effective signal priority strategies should balance the benefits for the transit vehicles with the negative impacts on the side traffic. Since the side-street volumes in the study network are not substantial to show marked changes in travel time for non-priority vehicles, three priority strategies (no priority, unconditional priority, load-based conditional priority) were also evaluated assuming a 40% increase in the demand entering the network from the side arterial Liljeholmsbron. Figure 5 summarizes the results with respect to average travel times. Load-based conditional priority strategies yield



**Advanced Public Transport Systems, Simulation-Based Evaluation. Figure 5**

Signal priority impact on travel times with increased side-street demand (source: [55])

similar benefits as unconditional priority. More importantly, the improvements in transit vehicle performance under conditional priority can be achieved with considerably less adverse impact on the side-street traffic. The results are consistent with those reported in a field study of schedule-based conditional priority in Eindhoven, the Netherlands [56]. The study, using data collected in the field, showed increases in side-street travel times of 100–300% at some intersections under unconditional priority, while the schedule-based conditional priority strategy produced significant reduction in transit travel times with only a minor impact on non-priority vehicles.

Automatic passenger counters (APC) may have measurement errors. For example, studies indicate that infrared APC systems on buses in Oregon have the tendency to overestimate passenger load, citing errors as high as 10% [60]. Thus, another scenario is simulated to determine whether passenger loads measured by APC with error might affect the performance of load-based conditional priority. Assuming that the APC count varies between 5% below and 15% above the true load, the rate of priority (% of buses receiving priority) increases from 66% (with no error) to 68% (with error). BRT buses experience travel times which on the average are slightly lower than with the same load-based strategy assuming no APC error. Furthermore, there was no effect on the side-street travel times. Hence, the strategy appears robust even under the assumed measurement errors. This analysis illustrates

another useful application of simulation. A simulation model, at the level of detail described in this paper, can be used to refine APTS strategies by explicitly modeling all aspects and characteristics of the technologies involved (including operating characteristics, such as measurements errors).

### Mesoscopic Transit Simulation Example: BusMEZZO

Mesoscopic transit simulation models are more appropriate for the analysis of the operations of transit systems at the network level. BusMEZZO is an example of such a model. The transit simulation component is directly integrated within the platform of Mezzo, a mesoscopic traffic simulation model [48]. Mezzo is an event-based simulator, which models vehicles individually. Links are divided into two parts: a running part, which contains vehicles that are not delayed by the downstream capacity limit, and a queuing part, which extends upstream from the end of the link when capacity is exceeded. The boundaries between the running and queuing parts are dynamic and depend on the extent of the queue. Vehicles have travel behavior and make route-choice decisions based on prior experience, available traffic information, etc.

### Simulating Transit Operations

Mesoscopic transit simulation requires similar functionalities as microscopic models. However, the level



of detail in the representation of some components, especially on the supply side, is lower.

**Supply Representation** The supply representation includes the transit network, the movement of vehicles and interactions with other traffic, the representation of the surveillance system, and the functionality of the transit operations control center.

Transit network, schedules, and runs are defined in a similar fashion as in the microscopic model presented in the previous section. However, vehicle movements and interactions are represented in simplified way. Buses enter the network according to the provided schedule, interact with the car traffic, and move according to prevailing conditions on the link (based on speed-density relationships). They stop at stops according to a dwell time that is calculated using models similar to the ones in microscopic simulation. Buses depart from stops according to the calculated dwell time or other considerations, such as timetable constraints for the specific route, or real-time control strategies that are implemented to improve the reliability of the route. Upon arrival at the end terminal, a bus either enters the next trip in its schedule or exits the network. With this process, delays encountered by buses on one trip influence its ability to perform the next trip on time. Hence, the model captures the propagation of delays along the route and is capable of producing the occurrence of bunching.

An important aspect in modeling transit supply is the representation of APTS both in terms of the sensor technology in place to collect data and the operating strategies. As in microscopic simulation, the model keeps track of data routinely collected by the APTS in place (AVL, APC, etc.). This data is processed in combination with schedule information preceding and following vehicles' data in order to generate information about the bus performance (e.g., schedule adherence and actual headways). This information, in turn, is used to trigger real-time strategies in order to improve operations. The model is capable of simulating a wide range of transit operations control strategies so that they can be evaluated and refined.

Strategies that are used to improve transit service reliability are of particular importance. Service reliability is often connected to headway variability. Headway variability impacts a number of important measures of

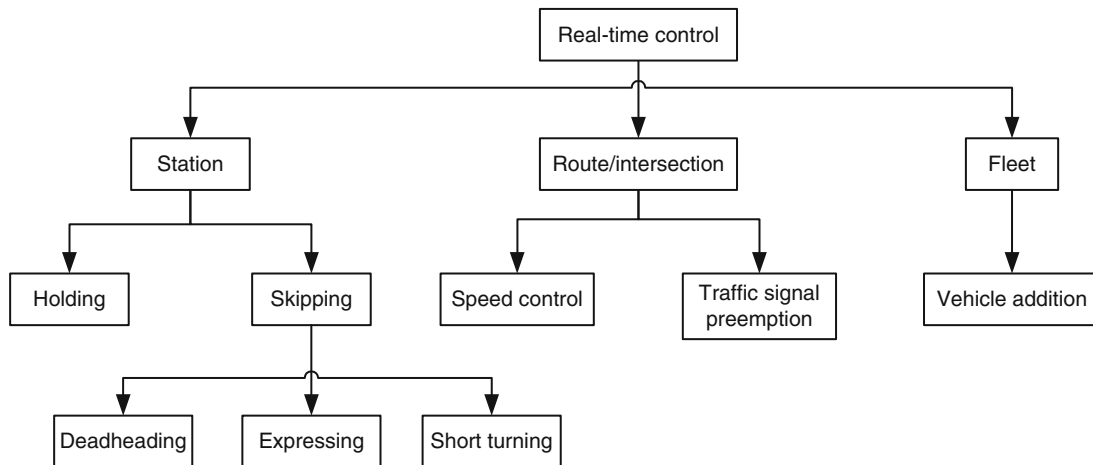
level of service for the passengers, such as waiting times, uncertainty with respect to arrival times at destinations, and passenger loads. Reliability also impacts utilization of resources. In this context, the model can be used to both assess the causes of unreliable service and develop strategies for effectively dealing with it. Service reliability has been the subject of many studies, including the comprehensive work reported in [61, 62]. A number of studies also exist that investigate the manifestation of service reliability using data collected through APTS technologies (e.g., [63–65]).

Real-time operations control has always been used as means to improve service reliability. Figure 6 illustrates a taxonomy of the various types of real-time control strategies [66]. A strategy may be applied at the station, route/intersection, or at the fleet level. Station-based strategies are either holding (vehicles held at stops for service restoration) or skipping strategies (where stations may be temporarily skipped during a particular trip). Skipping stops may take several forms, such as deadheading or expressing.

Holding is a particularly popular strategy because of the relative ease of implementation. A number of theoretical models have appeared that cover various aspects of the problem and apply them to both real and idealized scenarios. Originally, most of the work focused on simple network configurations with limiting assumptions and simplifications to make the problem tractable. Early research attempted to provide an understanding of the impact of holding strategies on passenger waiting times, with classic papers such as [67–69] leading the way in this direction. Recent work in the area includes [66, 70–73] and focuses on both evaluation of impacts and development of methods to design optimal strategies. Real-time control of buses for schedule coordination at terminals is studied in [74].

Because of the difficulties in obtaining analytical solutions to the problem, many studies rely on simulation to evaluate the developed strategies. For example, a simulation model of a single direction bus route was used to test several strategies, including holding, skipping, loading (but not unloading) passengers, and signal priority in [74]. Similarly, in [75], simulation was used to evaluate different real-time control strategies. BusMEZZO provides a platform where existing





**Advanced Public Transport Systems, Simulation-Based Evaluation. Figure 6**  
Real-time operations control strategies (source: [66])

strategies can be evaluated, and new approaches developed.

**Demand Representation** Passenger demand is modeled at various levels of detail (more information can be found in [76]):

- Arrival and alighting rates as discussed in section “Transit System Simulation.”
- Passenger origin–destination flows and related travel choices and travel behavior. The model can simulate individual passengers with attributes (e.g., age, income), origins, and destinations (assuming that the required information is available). A demand simulator component has been developed to deal with the diverse requirements of transit demand representation. Passengers are explicitly represented. Their behavior is captured dynamically as they make station and route (bus line) decisions based on experience, real-time transit information, and other service attributes. As mentioned in the previous section, in the context of mesoscopic models, transit assignment functionally is an important requirement. A dynamic transit assignment component is implemented based on two submodels: choice-set generation and path choice process. The generation of the choice set aims at reproducing the set of alternatives considered by passengers. The choice-set generation model is based on the transit network attributes –

transit routes, timetables, transferring stops and links, as well as access and egress links. However, in order to narrow the set of all available paths to a reasonable choice set, a number of filtering rules are used which apply logical and behavioral constraints.

Once the set of alternative paths has been generated, the path choice model determines how passengers choose between their alternatives. The path choice model takes into account traveler preferences and decision rules. A dynamic representation of the transit system facilitates the explicit modeling of capacity constraints. Moreover, passenger travel experiences and real-time information are important inputs to the path choice decisions. The dynamic assignment model considers the path choice as a decision process that consists of a sequence of dependent decisions. In general, there are two distinctive types of decisions: (a) stop-related decisions (origin stop choice, transfer stop choice, and destination stop choice) and (b) route-related decisions (choice among alternative transit routes). A discrete choice framework is used to model the various decisions [77].

**Measures of Effectiveness** Outputs from the simulation include stop level statistics, such as early and late arrivals, dwell times, numbers of boarding and alighting, bus loads, and travel times between stops. Aggregations at the level of the trip, the vehicle, or the line,

such as schedule adherence, headway and passenger wait time distributions, load profiles, and other level of service measures, are also computed.

### Case Study: Evaluation of Real-Time Holding Strategies

The mesoscopic model described in the previous section is used to evaluate holding strategies to improve service reliability of a busy urban route in Tel-Aviv, Israel (for more information see [78]). This high-demand route connects a dense suburb to the CBD. The scheduled headway during the peak period is 8 min, and the average running time is 49 min inbound and 41 min outbound.

The study compares the base case of no holding to two scenarios with holding strategies. In designing holding strategies, the main aspects of interest are the amount of time the bus is held, and the location of the stops (time points) where holding may be applied. In general, holding strategies are either schedule-based or headway-based. Schedule-based strategies apply control in order to ensure that buses run according to the timetable. Headway-based strategies aim at improving service by regulating the headways at stops.

Headway-based control is more appropriate for lines that have very frequent service (since in those cases, according to [54], passengers arrive at the stop randomly and do not plan their arrival according to schedule). Schedule-based strategies are more appropriate for low frequency routes with passengers coordinating their arrival with the schedule. Another important difference between the strategies is the amount of information required for their implementation. Schedule-based strategies require minimum information (a static timetable), while headway-based strategies require real-time information (that may vary depending on the exact nature of the strategy). In general, it is expected that more information should help the system perform better. A comparison between holding control strategies that rely only on local information to APTS-based strategies that incorporate more information was conducted in [75] in the context of coordination at transfer points. Not surprisingly, the study concluded that the best holding control strategy was the one that used all available information, including predicted arrival times of connecting buses,

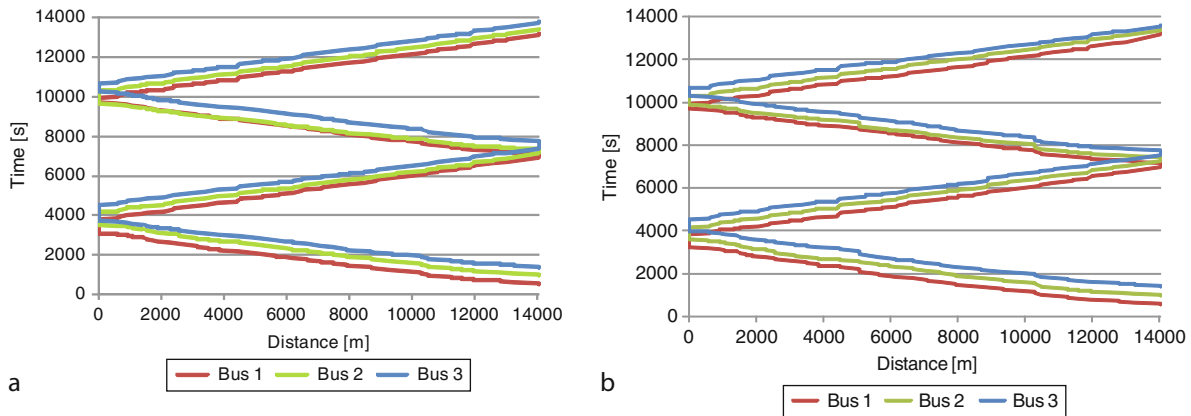
number of transferring passengers, and expected number of boarding passengers at downstream stops.

Schedule-based strategies hold buses that arrive early at a time point until a predefined time from their scheduled departure time is reached. Depending on the available information, more or less sophisticated headway-based control strategies can be implemented. For example, if only information regarding the preceding bus is available, the strategy takes the form of holding a bus until a certain threshold with respect to the scheduled headway is reached. If information is available with respect to both the preceding and following buses, then the holding strategy may take into account the headway with both buses and hold the subject vehicle until headways between them are equal (mean-headway strategy) or until a maximum holding time is reached [79, 80].

Regarding the location of the time points, strategies are typically applied at a subset of stops along a route [81]. Several studies (e.g., [82–85]) suggest that time-point stops should be located at the beginning of a sequence of high-demand stops. In contrast, [72] found that there are no significant benefits from defining time points along bus routes, except the origin terminal.

Following the recommendations in the literature, the design of the holding strategies that were tested had the following characteristics. The time points were located at three stops upstream of a sequence of high-demand stops. The holding time at each time was based on two strategies: a strategy that aims at equalizing the headway (when possible) between the subject bus and the preceding and following buses (mean headway), and a schedule-based holding. Therefore, the two strategies represent two rather opposite cases of information availability: minimum information (timetable) and relative high (arrivals of preceding and following buses at the time point).

Figure 7 illustrates the impact of control using a time-space diagram of the trajectories of three buses, each performing four chained trips. Sloped segments represent running time between stops, and vertical segments indicate dwell times and holding times at time points. Recovery times between trips at both terminals are also shown. Figure 7a illustrates the base case of no control, while Fig. 7b corresponds to the case of mean-headway control. Buses are experiencing



**Advanced Public Transport Systems, Simulation-Based Evaluation. Figure 7**

Time-space diagram of buses with no control (a) and mean-headway control (b) (source: [78])

bunching when there is no control, while the holding strategy is effective in preventing this by creating even headways at time points. For example, on the third trip, bus 2 is getting increasingly bunched with bus 1 under (a), a situation that is largely remedied when holding is applied.

At the system level, several measures of performance were calculated for each scenario, including average passenger waiting time, travel-time, and on-time performance. The schedule-based strategy resulted in very small benefits with respect to average waiting time for the passengers. In general, it had practically the same performance as the base case of no control with respect to all measures. The reason for this is that, with the timetable used for the route, only a small number of trips were ahead of schedule, and therefore the schedule-based strategy was used very infrequently. The mean-headway strategy reduced headway variability, as measured by the coefficient of variation, by as much as 30%. The average waiting time decreased by 5%. However, the mean-headway strategy results in service regularity gains at the expense of higher travel times and deterioration of on-time performance.

## Future Directions

Transportation systems are highly complex, and researchers, planners, and engineers often use simulation to study their performance, analyze and design possible system improvements, and develop decision

support systems to aid with their management. Over the past several years (motivated in part by the introduction of Intelligent Transportation Systems and APTS), transit simulation models have seen a lot of advances. As technology and policy innovations are constantly introduced, the required functionalities of transit simulation models also need to be expanded to satisfy the diverse needs of operations, planning, and policy evaluation. Information technology, for example, has created and will continue to create many opportunities for transit systems. These opportunities are twofold: (a) advances in sensor technologies allow operators to collect a lot of data and information about the system and (b) advances in communications technologies facilitate the dissemination of useful information to passengers, even personalized information on demand. In order to provide the functionality needed to support the new application foci, developments in the context of both microscopic models and mesoscopic models are needed.

On the microscopic side, seamless integration of traffic and transit driving behavior (at the individual vehicle level) needs to be further developed to reliably model APTS at the operational level. On the mesoscopic side, transit simulation needs to be closely integrated with developments in the area of simulation-based dynamic traffic assignment. Integrated, multimodal models that represent demand and supply at a consistent, disaggregate level, and capture the demand–supply interactions are important for the evaluation of future strategies to deal with sustainable

transportation systems. Multimodal models need to be sensitive to changes in the system introduced by policy and technological innovations. In this respect, better integration of demand and supply representations at a consistent level of detail is important and should be organized in modules with the following characteristics: (a) a demand module with disaggregate, dynamic, activity-based representation of the choices of individual travelers; (b) a supply module, with detailed, disaggregate, dynamic modeling of multimodal network performance through vehicle-based traffic simulation; and (c) a demand–supply interactions module with integration between the demand and supply modules through appropriate algorithms.

Integrated, simulation-based, dynamic, and disaggregate models have the potential to provide increased realism in capturing both travel behavior and network performance and so provide considerable flexibility in representing a wide range of systems and predicting their impacts. A number of challenges, both theoretical and practical, need to be resolved before such models become operational. Important challenges include the development of appropriate demand–supply algorithms that are able to predict the resulting behavior of the system, data and approaches for development and estimation of the constituent models, and the systematic validation and calibration of the overall model, and approaches to deal with the computational requirements of such large-scale models (more information about such a disaggregate, dynamic multimodal modeling of transportation systems can be found in [86]).

## Bibliography

### Primary Literature

1. Shannon RE (1975) *Systems simulation: the Art and science*. Prentice-Hall, Englewood Cliffs
2. United Nations Population Fund (2010) State of world population 2007. <http://www.unfpa.org/swp/2007/english/introduction.html>
3. Miles J, Chen K (2004) *ITS handbook*, 2nd edn. PIARC, Paris
4. Casey RF, Labell LN, Moniz L et al (2000) *Advanced public transportation systems – the state of the art: update 2000*. U.S. Department of Transportation, Federal Transit Administration (FTA), Washington, DC
5. Hwang M, Kemp J, Lerner-Lam E, Neuerburg N, Okunief P (2006) *Advanced public transportation systems: state of the art update 2006*, FTA-NJ-26-7062-06.1. Federal Transit Administration (FTA), Washington, DC
6. Furth P, Hemily B, Muller T, Strahman J (2006) Using archived AVL-APC data to improve transit performance and management, TCRP report 113. Transportation Research Board, Washington, DC
7. Furth P (2000) Data analysis for bus planning and monitoring, TCRP synthesis 34. Transportation Research Board, Washington, DC
8. Smith H, Hemily B, Ivanovic M (2006) *Transit signal priority (TSP): a planning and implementation handbook*. ITS America and US Department of Transport, Washington, DC
9. Hickman MD, Wilson NHM (1995) Passenger travel time and path choice implications of real-time transit information. *Transp Res C* 3(4):211–226
10. Nuzzolo A, Russo F, Crisalli U (2001) A doubly dynamic schedule-based assignment model for transit networks. *Transp Sci* 35(3):268–285
11. Wilson NHM, Zhao J, Rahbee A (2004) The potential impact of automated data collection systems on urban public transport planning. In: Wilson NHM, Nuzzolo A (eds) *Schedule-based dynamic transit modeling: theory and applications*. Academic, Dordrecht, pp 74–99
12. Vandebona U, Richardson AJ (1985) The effects of fare-collection strategies on transit level of service. *Transp Res Rec* 1036:79–87
13. Transportation Demand Management User Service (2005) National ITS architecture. US Department of Transportation, Washington, DC. [www.iteris.com/itsarch/html/user/usr18.htm](http://www.iteris.com/itsarch/html/user/usr18.htm)
14. Barceló J (ed) (2010) *Fundamentals of traffic simulation*, vol 145, International series in operations research & management science. Springer, New York
15. Daganzo C (1994) The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res B* 28(4):269–287
16. Messmer A, Papageorgiou M (1990) METANET: a macroscopic simulation program for motorway networks. *Traffic Eng Control* 31(9):466–470
17. Ben-Akiva M, Bierlaire M, Burton D, Koutsopoulos HN, Mishalani R (2001) Network state estimation and prediction for real-time transportation management applications. *Netw Spat Econ* 1(3/4):293–318
18. Ben-Akiva M, Koutsopoulos HN, Antoniou C, Balakrishna R (2010) Traffic simulation with DynaMIT (chap 10). In: Barceló J (ed) *Fundamentals of traffic simulation*. Springer, New York
19. Mahmassani HS (2001) Dynamic network traffic assignment and simulation methodology for advanced systems management applications. *Netw Spat Econ* 1(3/4):267–292
20. Yang Q, Koutsopoulos HN (1996) A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transp Res C* 4:113–129
21. Ben-Akiva M, Koutsopoulos HN, Toledo T, Yang Q, Choudhury C, Antoniou C, Balakrishna R (2010) Traffic simulation with MITSIMLab (chap 6). In: Barceló J (ed) *Fundamentals of traffic simulation*. Springer, New York

22. Burghout W, Koutsopoulos HN, Andreasson I (2005) Hybrid mesoscopic-microscopic traffic simulation. *Transp Res Rec* 1934:218–225
23. Algers S, Bernauer E, Boero M, Breheret L, Dougherty M, Fox K, Gabard JF (1997) Review of microsimulation models. In: SMARTTEST project deliverable D3, EU, Institute for Transport Studies, University of Leeds, Leeds. Available at: <http://www.its.leeds.ac.uk/projects/smartest/deliv3f.html>
24. Silva PCM (2001) Modelling interactions between bus operations and traffic flow. Doctoral dissertation, University College London
25. Khasnabis S, Karanti RR, Rudraraju RK (1996) NETSIM-based approach to evaluation of bus preemption strategies. *Transp Res Rec* 1554:80–89
26. Liu R, Clark S, Montgomery F, Watling D (1999) Microscopic modelling of traffic management measures for guided bus operation. In: Proceedings of eighth world conference on transport research, Antwerp
27. Chang J, Collura J, Dion F, Rakha H (2003) Evaluation of service reliability impacts of traffic signal priority strategies for bus transit. *Transp Res Rec* 1841:23–31
28. Lee J, Shalaby A, Greenough J, Bowie M, Hung S (2005) Advanced transit signal priority control using on-line microsimulation-based transit prediction model. In: Preprints of the transportation research board 84th annual meeting, Washington, DC
29. Ding Y, Chien S, Zayas A (2001) Simulating bus operations with enhanced corridor simulator. *Transp Res Rec* 1731:104–111
30. Morgan DJ (2002) A microscopic simulation laboratory for advanced public transport system evaluation. Master thesis, Massachusetts Institute of Technology
31. Liu CL, Pai TW (2006) Methods for path and service planning under route constraints. *Int J Comput Appl Technol* 25 (1):40–49
32. Liu R, Van Vliet D, Watling D (2006) Microsimulation models incorporating both demand and supply dynamics. *Transp Res A* 40:125–150, Washington, DC
33. <http://www.caliper.com/>
34. <http://www.ptvag.com/index.php?id=345>
35. BRT. <http://www.cfte.org/trends/brt.asp>
36. Werf JV (2004) SmartBRT: a tool for simulating, visualizing and evaluating bus rapid transit systems. California PATH Research Report, Institute of Transportation Studies, University of California, Berkeley
37. <http://www.paramics.com/>
38. Cortes CE, Fernandez R, Burgos V (2007) Modeling passengers, buses and stops in traffic microsimulators. The MISTRANSIT approach on the PARAMICS platform. In: Proceedings of the 86th transportation research board annual meeting, Washington, DC
39. Fernandez R, Cortes CE, Burgos V (2010) Microscopic simulation of transit operations: policy studies with the MISTRANSIT application programming interface. *Transp Plann Technol* 33(2):157–176
40. Wahba M, Shalaby A (2006) A general multi-agent modelling framework for the transit assignment problem – a learning-based approach. In: Innovative internet community systems. Springer, Berlin, pp 276–295
41. Wahba M, Shalaby A (2006) MILATRAS: a microsimulation platform for testing transit-ITS policies and technologies. In: Proceedings of the IEEE ITSC, Canada
42. Wang J, Wahba M, Miller EJ (2010) A comparison of an agent-based transit assignment procedure (MILATRAS) with conventional approaches. In: Proceedings of the 89th transportation research board annual meeting, Washington, DC
43. Boxill SA, Yu L (2000) An evaluation of traffic simulation models for supporting ITS development. Center for Transportation Training and Research, Texas Southern University
44. Abdelghany KF, Abdelghany AF, Mahmassani HS, Abdelfatah AS (2006) Modeling bus priority using intermodal dynamic network assignment-simulation methodology. *J Public Transp* 9(5):1–22
45. Meignan D, Simonin O, Koukam A (2007) Simulation and evaluation of urban bus-networks using a multiagent approach. *Simul Model Pract Theory* 15:659–671
46. Cats O, Toledo T, Burghout W, Koutsopoulos HN (2010) Mesoscopic modeling of bus public transportation. In: Proceedings and CD of the 89th transportation research board (TRB) annual meeting, Washington, DC
47. Toledo T, Cats O, Burghout W, Koutsopoulos HN (2010) Mesoscopic simulation for transit operations. *Transp Res C* 18(6):896–908. doi:10.1016/j.trc.2010.02.008
48. Burghout W (2004) Hybrid microscopic-mesoscopic traffic simulation. Doctoral dissertation, Royal Institute of Technology, Stockholm
49. Yang Q (1997) A simulation laboratory for evaluation of dynamic traffic management systems. Ph.D. thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
50. Yang Q, Koutsopoulos HN, Ben-Akiva M (2000) A simulation laboratory for evaluating dynamic traffic management systems. *Transp Res Rec* 1710:122–130
51. Levinson HS (1983) Analyzing transit travel time performance. *Transp Res Rec* 915:1–6
52. Lin T, Wilson NHM (1992) Dwell time relationships for light rail systems. *Transp Res Rec* 1361:287–295
53. TCRP Report 100 (2003) Transit capacity and quality of service manual (TCQSM), 2nd edn. Transportation Research Board, Washington, DC
54. Jolliffe JK, Hutchinson TP (1975) A behavioral explanation of the association between Bus and passenger arrivals at a bus stop. *Transp Sci* 9(3):248–282
55. Morgan D, Koutsopoulos HN, Ben-Akiva M (2003) Simulation-based evaluation of advanced public transportation systems. In: Wilson NHM, Nuzzolo A (eds) Schedule-based dynamic traffic modeling: theory and applications. Series in operations research/computer science interfaces, vol 28, chap 6. Academic, Boston, pp 95–102
56. Furth PG, Muller THJ (2000) Conditional bus priority at signalized intersections: better service quality with less traffic disruption. *Transp Res Rec* 1731:23–30



57. Chang J, Collura J, Dion F, Rakha H (2003) Evaluation of service reliability impacts of traffic signal priority strategies for bus transit. *Transp Res Rec* 1841:23–31
58. Kim W, Rilett LR (2005) An improved transit signal priority system for networks with nearside bus stops. *Transp Res Rec* 1925:205–214
59. Lee J, Shalaby A, Greenough J, Bowie M, Hung S (2005) Advanced transit signal priority control using on-line microsimulation based transit prediction model. *Transp Res Rec* 1925:185–194
60. Kimpel T, Strathman J (Apr 2002) Automatic passenger counter evaluation: implications for national transit database reporting. <http://www.upa.pdx.edu/CUS/PUBS/docs/PR124.pdf>
61. Abkowitz M, Slavin H, Waksman R, Englisler L, Wilson N (1978) Transit service reliability. Report UMTA-MA-06-0049-78-1. USDOT Transportation Systems Center, Cambridge
62. Abkowitz M, Tozzi J (1987) Research contributions to managing transit service reliability. *Adv Transp J* 21:47–65
63. Strathman JG, Dueker KJ, Kimpel T, Gerhart R, Turner K, Taylor P, Callas S, Griffin D, Hopper J (1999) Automated bus dispatching, operations control and service reliability. *Transp Res Rec* 1666:28–36, Washington, DC
64. Strathman JG, Kimpel TJ, Dueker KJ, Gerhart RL, Turner K, Griffin D, Callas S (2001) Bus transit operations control: review and experience involving Tri-Met's automated bus dispatching system. In: Transportation Research Board, 80th Annual Meeting, Washington, DC
65. Strathman JG, Kimpel TJ, Dueker KJ, Gerhart RL, Callas S (2002) Evaluation of transit operations: data applications of Tri-Met's automated bus dispatching system. *Transportation* 29:321–345
66. Eberlein XJ (1995) Real time strategies in transit operations: models and analysis. Ph.D. dissertation, Department of Civil Engineering, MIT
67. Osuna EE, Newell GF (1972) Control strategies for an idealized public transportation system. *Transp Sci* 6:52–72
68. Barnett A (1974) On controlling randomness in transit operations. *Transp Sci* 8(2):102–116
69. Barnett A (1978) Control strategies for transport systems with nonlinear waiting costs. *Transp Sci* 12(2):119–136
70. O'Dell S, Wilson N (1997) Optimal real-time control strategies for rail transit operations during disruption. In: Wilson N (ed) *Computer aided transit scheduling*, vol 471, Lecture note in economics and mathematical systems. Springer, Berlin, pp 299–323
71. Hickman M (2001) An analytic stochastic model for the transit vehicle holding problem. *Transp Sci* 35(3):215–237
72. Eberlein XJ, Wilson NHM, Bernstein D (2001) The holding problem with real-time information available. *Transp Sci* 35(1):1–18
73. Shen S, Wilson NHM (2001) Optimal integrated real-time disruption control model for rail transit systems. In: Voss S, Daduna J (eds) *Computer-aided scheduling of public transport*, vol 505, Lecture notes in economics and mathematical systems. Springer, Berlin, pp 335–364
74. Koffman D (1978) A simulation study of alternative real-time Bus headway control strategies. *Transp Res Rec* 663:41–46
75. Dessouky M, Hall R, Zhang L, Singh A (2003) Real-time control of buses for schedule coordination at a terminal. *Transp Res A* 37:145–164
76. Cats O, Koutsopoulos HN, Burghout W, Toledo T (2011) Effect of real-time transit information on dynamic passenger path choice. *Transportation Research Record. J Transp Res B* (in press)
77. Ben-Akiva M, Lerman S (1985) *Discrete choice analysis. Theory and application to travel demand*. MIT Press, Cambridge, MA
78. Cats O, Burghout W, Toledo T, Koutsopoulos HN (2010) Evaluation of real-time holding strategies for improved bus service reliability. Submitted for presentation and publication at the 2010 13th international IEEE conference on intelligent transportation systems (ITSC 10), Madeira Island
79. Koutsopoulos HN, Wang Z (2007) Simulation of urban rail operations. *Transp Res Rec* 2006:84–91
80. Daganzo CF (2009) A headway-based approach to eliminate bus bunching: systematic analysis and comparisons. *Transp Res B* 43:913–921
81. Abkowitz M, Lepofsky M (1990) Implementing headway-based reliability control on transit routes. *J Transp Eng* 116(1):49–63
82. Liu G, Wirasinghe SC (2001) A simulation model of reliable schedule design for a fixed transit route. *J Adv Transp* 35(2):145–174
83. Abkowitz M, Engelstein I (1984) Methods for maintaining transit service regularity. *Transp Res Rec* 961:1–8
84. Turnquist MA, Blume SW (1980) Evaluating potential effectiveness of headway control strategies for transit systems. *Transp Res Rec* 746:25–29, Washington, DC
85. Wirasinghe SC, Liu G (1995) Determination of the number and locations of time points in transit schedule design – case of a single run. *Ann Oper Res* 60:161–191
86. Ben-Akiva M, Bottom J, Gao S, Koutsopoulos HN, Wen Y (2007) Towards disaggregate dynamic travel forecasting models. *Tsinghua Sci Technol* 12(2):115–130

## Books and Reviews

- Ceder A (2007) *Public transit planning and operation: theory, modeling and practice*. Elsevier, Burlington
- Chung E, Dumont A (eds) (2009) *Transport simulation: beyond traditional approaches*. EPFL Press, Lausanne
- Sussman J (2005) *Perspectives on intelligent transportation systems (ITS)*. Springer, New York
- Vuchik V (2005) *Urban transit: operations, planning, and economics*. Wiley, Hoboken
- Vuchik V (2007) *Urban transit systems and technology*. Wiley, Hoboken
- Wilson NHM, Nuzzolo A (eds) (2004) *Schedule-based dynamic transit modeling: theory and applications*. Academic, Dordrecht



## Aerosol in Global Atmosphere

COLIN O'DOWD

School of Physics & Centre for Climate and Air  
Pollution Studies, Ryan Institute, National University  
of Ireland Galway, Galway, Ireland

### Article Outline

Glossary

Definition of the Subject

Introduction: The Nature and Impacts of Atmospheric  
Aerosols

Aerosol Formation, Dimensions, and Removal

Global Aerosol Budgets, Burdens, and Spatial  
Distributions

Water Uptake by Aerosols

Aerosols, Global Radiative Budget, Precipitation, and  
Climate

Aerosols and Air Pollution

Future Directions: Interactions Between Aerosol Air  
Pollution and Climate

Bibliography

### Glossary

**Aerosol** A suspension of liquid or solid particles in a fluid (in this context air). Also referred to as particulate matter (PM).

**Homogeneous nucleation** Spontaneous formation of an aerosol particle from vapor phase precursors through a phase transition, without need of a particle nucleus.

**Hygroscopicity** The ability to absorb water from the atmosphere.

**Cloud droplet activation** Spontaneous condensation growth of a wettable aerosol particle into a cloud droplet.

**Aerosol optical depth** Quantitative measure of the extinction of solar radiation by aerosol scattering and absorption between the point of observation and the top of the atmosphere.

**Radiative forcing** The radiative forcing of the surface-troposphere system due to the perturbation in or the introduction of an agent (say, a change in greenhouse gas concentration) is defined as the

change in net (down minus up) irradiance (solar plus longwave in  $\text{Wm}^{-2}$ ) at the tropopause.

**Climate** Mean values of meteorological parameters in a given region over relatively long periods of time.

### Definition of the Subject

Atmospheric aerosols are airborne liquid or solid particles, varying in size from nanometers to hundreds of microns. Atmospheric aerosols are formed from both natural and anthropogenic sources and can have complex chemical speciation and shape. They are formed from mechanical breakup of a parent material, such as dust or sea spray (so-called primary aerosol formation), or from gas-to-particle conversion processes, such as condensation of semivolatile organic compounds (VOCs) or chemical reactions in the aerosol phase such as oxidation of  $\text{SO}_2$  to  $\text{SO}_4$  (so-called secondary aerosol formation). These tiny airborne particles can form haze and cloud layers that influence the Earth's climate: aerosols directly reflect incoming solar radiation back out to space, providing a direct radiative effect on climate, some of which act as cloud condensation nuclei in the cloud formation process where they determine cloud microphysical and radiative properties, ultimately determining cloud reflectance providing an indirect radiative effect. The net climatic impact of atmospheric aerosols is to provide a cooling effect that partially offsets greenhouse gas-induced global warming. Aerosols also contribute to poor air quality and adverse health effects, typically resulting in reduced life span and increased mortality rates in the most polluted environments.

### Introduction: The Nature and Impacts of Atmospheric Aerosols

Atmospheric aerosols are tiny airborne liquid or solid droplets or particles, ranging from nanometers to tens or even hundreds of microns in size. They are diverse in shape and chemical composition and can possess varying optical properties and hygroscopic (water uptake) properties. Particle shape ranges from the most simple to characterize, that is, spherical, to complex fractal agglomerates with similar diversity in aerosol density. Chemical composition ranges from relatively simple inorganic composition to very complex organic matter

composition and mixtures thereof. The vast majority of organic chemical species have not, to date, even been identified. The aerosol population is a dynamic physical and chemical evolving system with fresh aerosol plumes often being externally mixed in terms of chemical composition, evolving with time into more complex internal mixtures. Primary aerosol production is the breakup of a parent material or incomplete combustion and results in the direct injection of aerosol particles into the atmosphere. Secondary aerosol production is the formation of aerosol phase products from nucleation, condensation, and multiphase chemical reactions resulting in aerosol production in the atmosphere itself as opposed to injection into the atmosphere. Aerosol particles in the size range from 100 nm to 1  $\mu$ m are particularly efficient at scattering, and depending on the chemical composition, absorption of solar radiation. Sea-salt, sulfate, and nitrate are all efficient scatters, while soot or black carbon and dust are efficient absorbers. These important optical properties lead to a reduction of visibility from aerosol haze layers under conditions of high aerosol concentrations and such haze layers are responsible for a number of atmospheric optical phenomena such as beautiful sunsets often seen under clear sky conditions. Aerosol particles also act as condensation nuclei for cloud droplets to form on. Without condensation nuclei, it would be virtually impossible for clouds to form, therefore aerosols also play an important role in the hydrological cycle. Cloud drops are also efficient scatters of solar radiation and consequently, both aerosol haze and cloud layers play an important role in the global radiative budget through contributing to 50% or more of the Earth's albedo. The significant contribution to global albedo necessitates an important role in climate, and in terms of a changing global aerosol population, a role in climate change also exists. Aerosols are both natural and anthropogenic, the former produced by the biosphere (sulfate and organics) and wind interaction at the Earth's surface (sea spray and dust production), the latter produced by industrial, transport, and combustion emissions (sulfate, nitrate, soot, and organics). Biosphere emissions are connected to climate change feedbacks and virtually impossible to control while anthropogenic emissions can be controlled to a certain extent. Although some aerosol species absorb solar radiation, partially contributing to

global warming, the predominant climate effect is one of net global cooling. Since the start of the industrial era, aerosol-related air pollution has, to a significant degree, been masking the global warming impact of greenhouse gases and has been slowing the rate of global warming. The effect on climate can be considered as a welcome impact of atmospheric aerosols; however, micron-sized and submicron-sized aerosol also have a high penetration efficiency through the human respiratory system leading to increased respiratory and cardiovascular health impacts. The adverse health impacts lead to increased mortality rates and reduced life span. For example, the London black smoke pollution events in the 1950s resulted in thousands of excess deaths over periods lasting the order of several days. To counteract the adverse health impacts of atmospheric aerosols, various clean air acts have been introduced in various continents, particularly developed industrial continents. Such legislation, while having the required successful impact on air quality and public health, has had a negative impact on global warming in that the reflective haze and cloud layers have become less reflective with the effect of reducing the masking effect on global warming. Particularly over the last two to three decades, cleaner air has resulted in a remarkable increase in the rate of global temperature increase as greenhouse gas concentrations in the atmosphere continue to rise. The coupling between air pollution, air quality, and climate change has led to the synergetic development of air pollution and climate change policy to achieve optimal effect in attempting to mitigate and abate climate change while simultaneously striving for clean air.

The challenge, however, is great in terms of producing accurate assessments of aerosol impacts, particularly those relating to climate change because the atmospheric aerosol system is incredibly complex and this complex system needs to be incorporated into large-scale and global climate assessment and prediction models. From the emission and formation stage of the aerosol life cycle, through its evolution, interaction with clouds and radiation, to its removal from the atmosphere is a highly complex non-linear system spanning spatial and temporal scales from nanometers and seconds (in terms of nucleation of new particles, to global and decadal scales, respectively, in terms of climatic impacts. The complexity is further compounded

by the diverse and even incomplete quantification of the chemical speciation of atmospheric aerosol particles. This article summarizes the nature, formation, transformation, distribution and impacts of aerosols in the global atmosphere and documents the recent advances in the knowledge relating to atmospheric aerosols and their impacts.

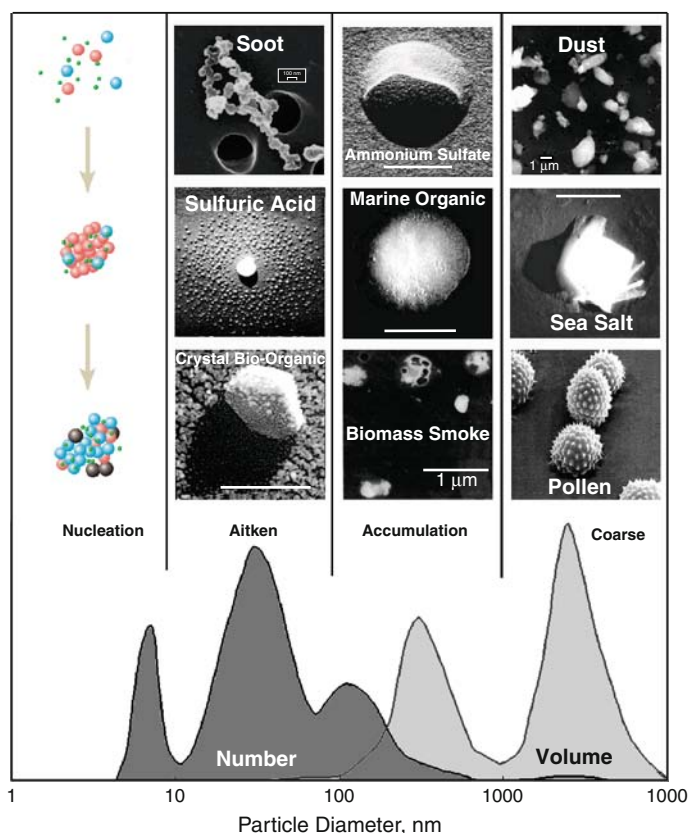
### Aerosol Formation, Dimensions, and Removal

Aerosols are produced via two generic mechanisms: (1) primary aerosol formation and (2) secondary aerosol formation. Primary aerosols are formed from the mechanical breakup of a parent material, with the resultant aerosol inheriting the chemical properties of the parent bulk material. In addition, primary aerosol production can result from incomplete combustion processes. These processes result in the direct emission of aerosols into the atmosphere. Secondary aerosols, on the other hand, are formed via gas-to-particle conversion processes such as oxidation or other means of chemical reduction of volatile precursor gases, condensation of low-volatility gases onto existing aerosol, homogeneous nucleation (phase-transition) producing new aerosol particles, aqueous phase reactions of dissolved gases producing stable aerosol phase species, or chemical reactions on aerosol surfaces leading to an increase in the condensed aerosol mass. In contrast to direct ejection of primary aerosols into the atmosphere, which necessitates a surface source for primary aerosol (excluding aviation emissions), secondary aerosols are formed in situ in the atmosphere, although many of their precursors are likely to originate at the surface. The formation of secondary aerosols can occur at all levels in the atmosphere, from the boundary layer in the lower troposphere to above the stratosphere.

Once formed, an aerosol particle may undergo physical and chemical transformation, but generally there are only two main sinks – that is through dry deposition at the surface, or wet deposition involving washout by precipitation. Dry deposition is the transport of aerosol particles from the atmosphere to the surface in the absence of precipitation. The level of turbulence in the atmosphere, particularly that close to the ground, determines the rate at which particles are deposited [1]. The deposition rate depends on

surface type (i.e., roughness) and particle size since there also is a gravitational settling component to dry deposition. Combining both wet and dry deposition removal processes, on average, submicron particles reside in the boundary layer for up to  $\sim 4\text{--}8$  days [2] depending on precipitation, while supermicron particles, depending on the size, have lifetimes from an hour to 1–2 days. For example, a 100-nm particle takes approximately 36 years to fall 1 km whereas a 10- $\mu\text{m}$  particle takes 3.6 days. Clearly, if particles are vertically transported to the free troposphere or the stratosphere, their lifetime in the atmosphere is greatly increased even to timescales of a year or more in the stratosphere. Removal in the atmospheric boundary layer is increased due to turbulent effects. In terms of aerosol-related air pollution, effects are more so on a local-to-regional scale rather than on a global scale due to removal timescales; however, as it will be seen later, in a climate context, a small amount of aerosol pollution transported to very remote regions can have a notable impact, despite relatively short atmospheric residence timescales. Physical transformations include condensation growth and coagulation (collisions of two particles resulting in one larger particle), while chemical transformations include changes in chemical composition from chemical reactions (e.g., oxidation), along with condensation of different vapors to that already in the aerosol phase. These transformations result in the aerosol population becoming “internally mixed” at long distances from source regions as opposed to “externally mixed” near source regions.

Primary production results typically in larger-micron- or supermicron-sized particles, while secondary production results in nanometer- to submicron-sized particles, leading to size distributions spanning many orders of magnitudes. In Fig. 1, a typical number size distribution is displayed. The peak in the number frequency distribution is in submicron sizes; however, for the same distribution, the volume or mass distribution is more often than not in supermicron sizes, even though the contribution of supermicron particles generally only comprises less than 1% of the total number concentration. Similarly, the mass of particles where the peak concentration occurs contribute less than 1% of the total mass. The peak in aerosol surface area (not shown) usually lies between the number concentration and volume peaks. The size



**Aerosol in Global Atmosphere. Figure 1**

A typical aerosol size distribution covering 4 orders of magnitude in size. Two moments of the size distribution are shown: number and volume. Also shown is a variety of different aerosol types from scanning electron microscope images (Copyright Springer. Modified from [3])

distribution comprises a number of “log-normal” modes reflecting different formation and transformation processes. At the smallest sizes is the nucleation mode where new particles, nanometers in size, are formed from nucleation of condensable vapors. As the nucleation mode grows through further condensation, an Aitken mode (called after the Scottish Scientist John Aitken) of a few tens of nanometers is formed. Through further transformations and growth processes, the Aitken mode grows into an accumulation mode of between about 100 and 200 nm. Generally, the coarse or supermicron mode arises from the direct injection of primary particles. The accumulation mode is termed so as removal processes are least efficient at these sizes and particles tend to accumulate there. Both smaller particles (in the nucleation and

Aitken modes) and larger (supermicron mode) particles have notably shorter lifetimes than that associated with accumulation mode particles but for different reasons. Smaller particles are highly diffusive and either self-coagulate to form a larger particle, or they coagulate with larger particles (with a high surface area) where they are effectively removed from the size distribution. Their mass is so small that gravitational settling is negligible. In contrast, larger particles have a negligible diffusion coefficient, but significant mass so that gravitational settling determines their lifetime. Also shown in Fig. 1 are scanning electron micrographs of various aerosol types such as soot (black) carbon, marine organic, sulfate, and bioaerosol particles. These images demonstrate the complexity of atmospheric aerosols in terms of shape, chemistry, and resultant

density, and this complexity is propagated through measurement and modeling challenges associated with quantification of aerosol properties and effects.

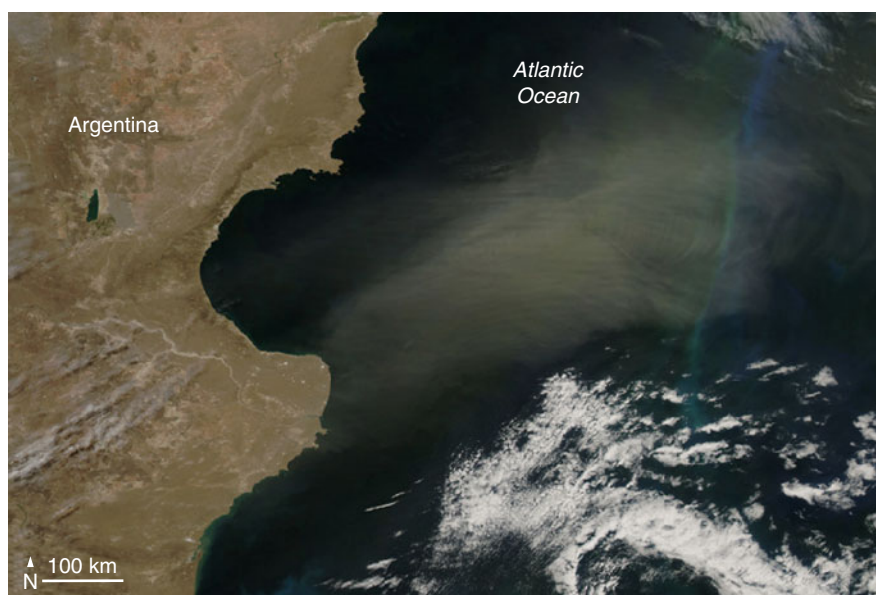
### Primary Aerosol Formation

In the atmosphere, the most important natural primary aerosol types result from the interaction of wind with terrestrial or oceanic surfaces. For terrestrial primary aerosols, wind interaction with agricultural and desert soils lifts loose fragments of the parent material through saltation and suspension processes resulting in airborne dust particles [4]. These primary dust particles are typically supermicron in size, comprising low number concentration, but high surface area and mass concentrations. The most dramatic example of this is the vertical uplifting of large amounts of desert dust into aerosol plumes that can extend for thousands of kilometers (Fig. 2). These plumes often are transported and deposited into oceanic regions where they can provide an important source of nutrients for the marine biota [5].

Wind interaction with terrestrial vegetation can also trigger the release and dispersion of bioaerosol such as pollen and seeds as well as waxlike and leaf-fragment debris. Humic matter, plant decomposition products, fungi, and microbes are also released from

plants. Fungi, in turn, release spores through a variety of ejection processes [6]. Primary bioaerosols also include entire organisms such as bacteria, fungi, and viruses as well as debris [7, 8].

Over oceanic water surfaces, wind stress at the air–sea interface leads to breaking waves and the entrainment of air into the water column, producing whitecaps [9, 10]. Whitecaps are ensembles of bubbles of entrained air rising to the surface where they burst due to the pressure differential inside and outside of the bubble. When the bubble film bursts, it breaks up into hundreds of individual seawater drops that rapidly equilibrate to ambient humidity leaving behind sea-salt “film” droplets, 0.01–1  $\mu\text{m}$  in size and comprising a concentrated solution of the ionic compounds found in sea-salt and water. After the film bursts, the cavity remaining collapses through instabilities leading to the formation of an upward movement of a water jet which fragments into a few jet droplets, typically 1–10  $\mu\text{m}$  in size, of similar composition to film drops [11]. Film and jet drops are produced at wind speeds greater than 3–4  $\text{m s}^{-1}$  which is observed to be the threshold for the onset of whitecaps. At wind speeds of 8–10  $\text{m s}^{-1}$ , the wind stress is sufficient to tear off the crests of waves and produce spume droplets, typically 10–100  $\mu\text{m}$  in size. The flux of sea spray follows a power-law



**Aerosol in Global Atmosphere. Figure 2**

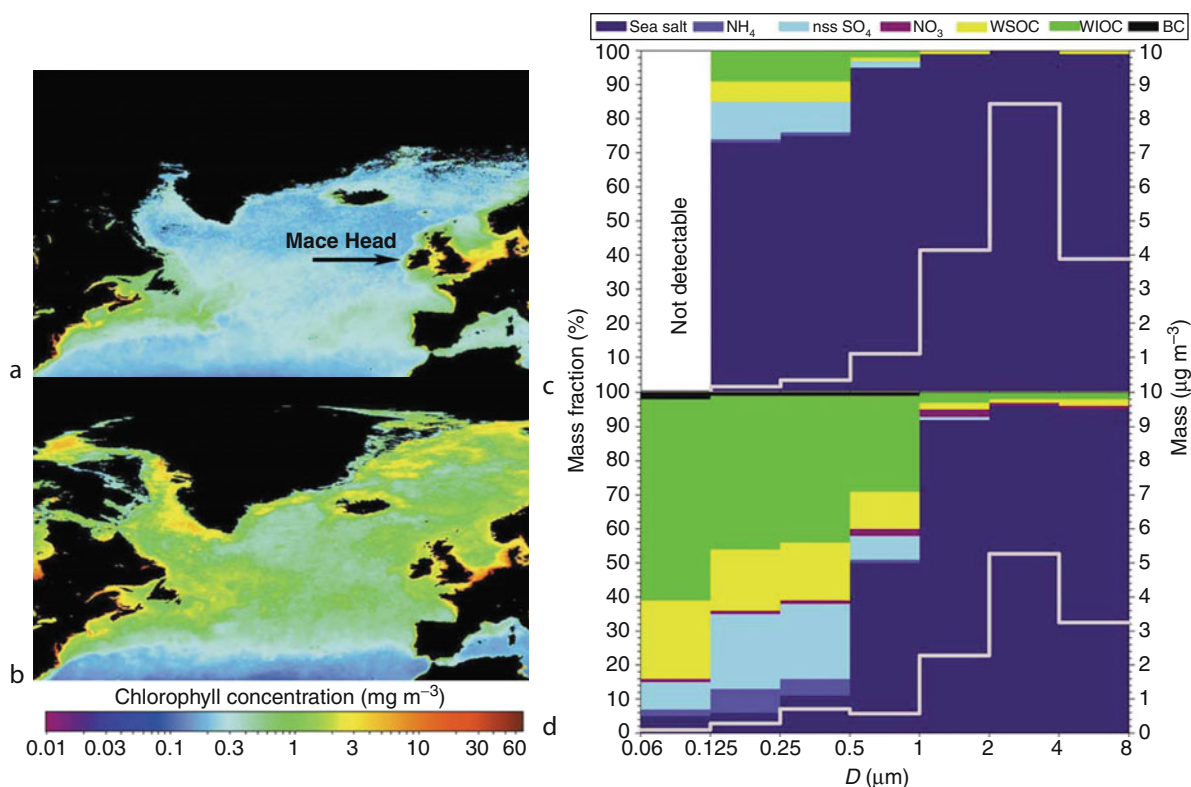
Patagonian dust over the Atlantic Ocean (Courtesy of NASA-MODIS)



dependency on wind speed [12]. In biologically active waters, organic matter tends to become concentrated at the ocean surface and at the bubble–water interface. This organic matter displaces seawater in the bubble film, ultimately displacing the ionic sea-salt in the film and jet drops produced, leading to the enrichment of organic matter in sea-spray aerosol [13]. The enrichment is observed to increase with decreasing particle size. The vast bulk of the organic matter enriched in sea-spray aerosol is water-insoluble organic matter. The chemical composition of marine aerosol is shown in Fig. 3 for periods of low and high biological activity over the North Atlantic as illustrated by oceanic

chlorophyll-*a* concentrations. Supermicron sizes contain predominantly sea-salt, while for submicron sizes, the enrichment of water-insoluble organic matter is linked to biological activity in the ocean (the water-soluble organic matter in the figure is also thought to be marine biogenic in origin but formed from secondary processes).

In terms of anthropogenic sources of primary aerosol, incomplete combustion (mainly from fossil fuel, but also from biomass burning and cooking) results in pyrogenic aerosols in the form of primary organic matter and soot (black) carbon particles. Fossil fuel emissions result from industrial, power plant, transport, and



**Aerosol in Global Atmosphere. Figure 3**

*Left:* Organic matter in sea-surface waters. SeaWiFS-derived seasonal average sea-surface chlorophyll-*a* concentrations in winter (a) and spring (b), illustrating low biological activity in North Atlantic waters during low and high biologically active periods (courtesy of SeaWiFS Project, NASA/Goddard Space Flight Center and ORBIMAGE). The location of Mace Head is shown in (a). Marine air masses arrive at Mace Head after at least 96 h transit over the ocean from the Arctic and northwest Atlantic. *Right:* Chemical composition of marine aerosols. Shown are average size-segregated chemical compositions and absolute mass concentrations for North Atlantic marine aerosols sampled with a Berner Impactor, for low biologically active periods (c) and high biologically active periods (d). The concentrations of water-soluble organic carbon, water-insoluble organic carbon, and black carbon are reported as mass of organic matter (Reprinted from [13])



residential heating sources and the resulting aerosol chemical composition varies according to the fuel type and the combustion technology used [14, 15]. Coal combustion results in the emission of supermicron fly-ash particles, while petroleum combustion results in predominantly submicron carbonaceous aerosols comprising soot carbon and organic material [16]. In addition to primary emissions, secondary aerosol precursors such as SO<sub>2</sub>, NO<sub>x</sub>, and hydrocarbons are also emitted from combustion sources.

### Secondary Aerosol Formation via Nucleation

Homogeneous nucleation is the term used for the spontaneous formation of an aerosol particle from vapor phase precursors through a phase transition. Homogeneous-homomolecular nucleation is the self-nucleation of a single species, while homogeneous-heteromolecular nucleation is the self-nucleation of two or more species. Homogeneous nucleation occurs in a supersaturated vapor phase where the saturation ratio ( $S$ ) of a species in air is the ratio of the species' partial pressure to the saturation vapor pressure in equilibrium with its liquid phase at a given temperature. The Classical Theory of nucleation is based on rate equations relating to the random formation of vapor clusters of more than one vapor molecule and the evaporation of these clusters [1]. The majority of these clusters evaporate as quickly as they form since they are generally thermodynamically unstable. If a sufficient number of the right low-vapor pressure molecules collide, the cluster can achieve a critical radius and achieve thermodynamic equilibrium (i.e., where the evaporation rate is zero). The nucleation rate is the net number of clusters per unit time that will grow past the critical size.

In addition to the kinetic aspect of forming clusters, the thermodynamics of the phase change has to be considered. In a thermodynamic process, such as a phase change resulting in chemical potential equilibrium of the molecular components, the change in the Gibbs free energy ( $G$ ) must be zero or negative. This condition must also hold for a curved surface of an aerosol particle or droplet. The curved surface will affect the equilibrium vapor pressure in a manner that a higher vapor pressure is required to maintain equilibrium for a smaller droplet (i.e., higher curvature effect). The change in Gibbs free energy  $\Delta G$  associated

with the formation of a single drop of a pure substance or radius  $R_p$  is given by:

$$\Delta G = \Delta G_{\text{droplet}} - \Delta G_{\text{pure\_vapor}}$$

which after expanding and rearranging, the right-hand side terms become:

$$\Delta G = -\frac{4}{3}\pi R_p^3 \frac{kT}{v_l} \ln S + 4\pi R_p^2 \sigma$$

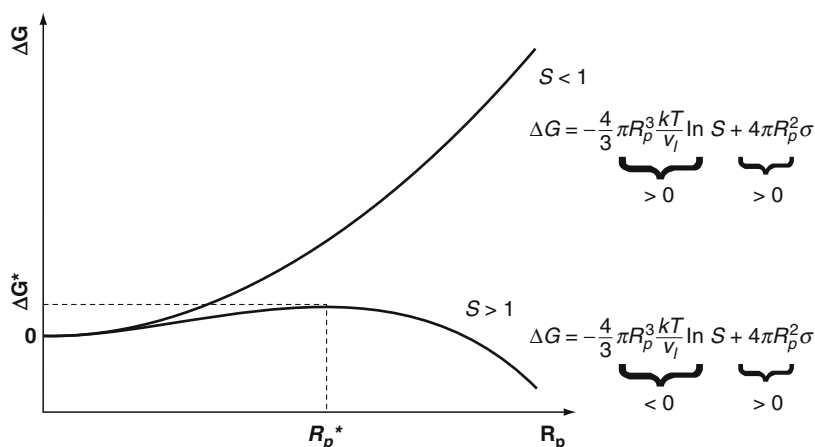
The second term on the right-hand side is the free energy associated with an interface with radius of curvature  $R_p$ .  $v_l$  is the volume occupied by a molecule in the liquid phase,  $S$  is the supersaturation,  $T$  is the temperature, and  $\sigma$  is the surface tension. The first term is a function of the change in free energy for a molecule going from the gas phase to the liquid phase (ultimately a function of the saturation ratio,  $S$ ) and the particle volume. It is seen, in Fig. 4, that if  $S < 1$ , both terms are positive and  $\Delta G$  increases monotonically. However, if  $S > 1$ ,  $\Delta G$  contains both positive and negative terms and for small droplet radii the second term dominates with  $\Delta G$  increasing as for the scenario for  $S < 1$ . As the radius increases, the first term (providing a reduction in  $\Delta G$ ) progressively becomes more important until  $\Delta G$  reaches a maximum at  $\Delta G^*$ , corresponding to a critical radius of  $R_p^*$ . Above  $R_p^*$ , the first term starts to dominate, resulting in a reduction in  $\Delta G$ . The critical radius can be determined by setting the derivative of  $\Delta G = 0$  and solving, resulting in:

$$R_p^* = \frac{2\sigma v_l}{kT \ln S}$$

$\Delta G^*$  represents a metastable equilibrium state where if one molecule evaporates from the droplet of critical radius, the droplet will continue to evaporate and shrink, while if one molecule is added to the droplet, the droplet will continue to grow.  $\Delta G^*$  represents a free energy barrier that the system has to overcome in order to grow and become stable.

Rearranging the previous equation gives the so-called Kelvin equation that describes the equilibrium vapor pressure of a vapor over a droplet relative to that over a flat surface:

$$S = \exp\left(\frac{2\sigma v_l}{kTR_p}\right) \quad p_A = p_A^0 \exp\left(\frac{2\sigma M}{RT\rho_1 R_p}\right)$$



**Aerosol in Global Atmosphere. Figure 4**

Gibbs free energy change for formation of a droplet of radius  $R_p$  from a vapor with saturation ratio  $S$

where  $M$  is the molecular weight of the substances and  $\rho_l$  is the liquid phase density.

The Kelvin equation demonstrates an exponentially increasing equilibrium vapor pressure with decreasing particle size. From this equation, a saturation ratio of the order of 10 is required for the smallest water vapor clusters to be thermodynamically stable and grow. The critical radius in this case comprises of the order of 20 water molecules. For realistic supersaturations encountered in the atmosphere (e.g.,  $S = 1.01\%$ ), the critical radius is  $0.12 \mu\text{m}$  and contains  $>10^8$  molecules. As a result, it is seen that water vapor does not participate in homogeneous nucleation in the atmosphere.

For homogeneous-homomolecular nucleation, supersaturation with respect to the species is required. In contrast, this is not the case for heteromolecular nucleation where nucleation can take place when the individual species are subsaturated with respect to the pure substances but supersaturated with respect to the resulting solution droplet containing these substances. Binary homogeneous nucleation is a process typically considered to occur in the atmosphere and the theory is an extension of homogeneous-homomolecular nucleation to account for two species.

Air contains many rapidly moving and colliding trace gases of varying vapor pressures. Those with low saturation vapor pressures, given the right conditions, prefer to be in the liquid solution phase or the solid phase from a thermodynamic perspective. The random

collisions of this soup of molecules result in the continuous formation of clusters of a few to tens of molecules. Such stable clusters will then grow in size with each subsequent collision of a low vapor pressure molecule. The size of a critical cluster is of the order of  $0.5 \text{ nm}$ ; however, at this size, the cluster has a very high diffusion coefficient and unless it grows rapidly, it will be scavenged by diffusion to larger, preexisting aerosol particles. For example, if the cluster can grow to  $3 \text{ nm}$ , its diffusion coefficient is reduced by a factor of 10 and if it grows to  $6 \text{ nm}$ , it is reduced further by a factor of 10 [17]. In other words, if the cluster can grow to  $6 \text{ nm}$ , its probability of survival is increased by a factor of 100.

In the atmosphere, very few vapors can participate in homogeneous nucleation. The best candidate is sulfuric acid, which in conjunction with water vapor can participate in binary homogeneous nucleation. It should be noted that the probability of stable cluster formation (or nucleation) is exponentially inversely dependent on temperature, and as a result, binary nucleation of sulfuric acid and water vapor is likely to occur only in the upper troposphere and the stratosphere where temperatures are sufficiently low to promote nucleation under atmospheric concentrations of sulfuric acid. In the troposphere, a different mechanism is required for nucleation to proceed under atmospheric concentrations of sulfuric acid concentrations (typically  $10^6$ – $10^7 \text{ molecules cm}^{-3}$ ). In the presence of ammonia, classical theory predicts that ternary nucleation of ammonia, sulfuric acid, and water can readily

proceed to more readily produce thermodynamically stable clusters under tropospheric conditions. The presence of ammonia lowers the thermodynamic barrier which has to be overcome to form an embryo. In classical theory, whether binary or ternary nucleation, the number of molecules in a critical cluster is of the order of 10 or more.

According to the nucleation theorem [18], the slope of the relationship between the nucleation rate and sulfuric acid concentration corresponds to the number of molecules in a critical cluster:

$$N_{\text{crit}} = d(\ln J)/d(\ln[\text{H}_2\text{SO}_4])$$

Atmospheric measurements of the nucleation rate versus sulfuric acid concentration reveal a slope of between 1 and 2 [19, 20] as do the most recent laboratory studies of nucleation. A slope of 2 can be explained by collision controlled or kinetic nucleation [21] while a slope of 1 can be explained by two processes: one being the requirement of an additional stabilizing or condensing vapor participating in the initial growth of the clusters; the other being activation of clusters by a supersaturated vapor [22, 23]. It is thought that the acid clusters are activated into aerosol particles in the presence of a supersaturated organic vapor field in the same way that cloud nuclei are activated into cloud droplets in a supersaturated water vapor field (the theory behind activation will be discussed later in the section on aerosol–cloud interactions). Recently, the formation of organosulfate clusters was suggested to explain the chemistry behind the cluster activation or kinetic mechanisms and thus atmospheric nucleation [24]. The aforementioned analysis reveals a critical cluster diameter of 0.8 nm and 1–2 sulfuric acid molecules in the cluster. The most recent laboratory studies [25] reveal that nucleation occurs at  $\text{H}_2\text{SO}_4$  concentrations similar to those found in the ambient atmosphere during nucleation events and that the measured particle formation rates are proportional to the product of the concentrations of  $\text{H}_2\text{SO}_4$  and an organic molecule. This suggests that only one  $\text{H}_2\text{SO}_4$  molecule and one organic molecule are involved in the rate-limiting step of the observed nucleation process.

Indirect measurements of 3–10 nm particle chemical composition immediately after a natural new particle production event over forested regions reveal that the chemical composition of newly formed particles are

not sulfuric acid or aerosol sulfate [26], but comprise more so of condensed organic matter. Calculations also reveal that there is insufficient sulfuric acid to nucleate new clusters and to grow these into aerosol particles larger than 3 nm. Whereas sulfuric acid appears to drive the nucleation of new clusters, an additional species is required to grow the clusters into aerosol particles. Under very polluted conditions, there is typically sufficient availability of sulfuric acid to both form stable clusters and to grow these clusters into aerosol particles tens to hundreds of nanometers in size. During particle production events in background air, the aerosol population typically increases from concentrations of  $\sim 1,000$  to  $\sim 100,000 \text{ cm}^{-3}$ , while in polluted air concentrations increase from  $\sim 10,000$  to  $\sim 100,000 \text{ cm}^{-3}$  [22, 23].

The most dramatic and intense nucleation events occur in coastal air where background concentrations increase from  $\sim 500$  to  $\sim 1,000,000 \text{ cm}^{-3}$  [26, 27]. These events occur when seaweed beds are exposed during low tide conditions. The algae emit  $\text{I}_2$  and  $\text{CH}_2\text{I}_2$  which are rapidly photolyzed and oxidized by  $\text{O}_3$  resulting in the formation of iodine oxides which readily nucleate into stable clusters. In addition, there is sufficient iodine oxide produced to account for the rapid growth of the clusters into aerosol particles of 100 nm [28].

### Multiphase Secondary Aerosol Production

The vast majority of secondary aerosol mass is not produced through the nucleation process but through condensation, aqueous phase reactions, and multiphase processes. Condensation typically involves both inorganic low-volatility gases, such as sulfuric acid and nitric acid, and a vast range of condensable organic species. Sulfuric acid results from the oxidation of  $\text{SO}_2$  by the OH radical, while nitric acid results from the oxidation of  $\text{NO}_2$  in the atmosphere.  $\text{NH}_3$  can also condense on acidic particles resulting in fully or partially neutralized particles. Nitric acid is regarded as a semivolatile condensable gas that can be thermodynamically stable in either the gas phase or the aerosol phase, depending on environmental conditions. At lower temperatures and higher relative humidities, nitric acid is more likely to be found in the condensed aerosol phase, while at warmer temperatures and lower

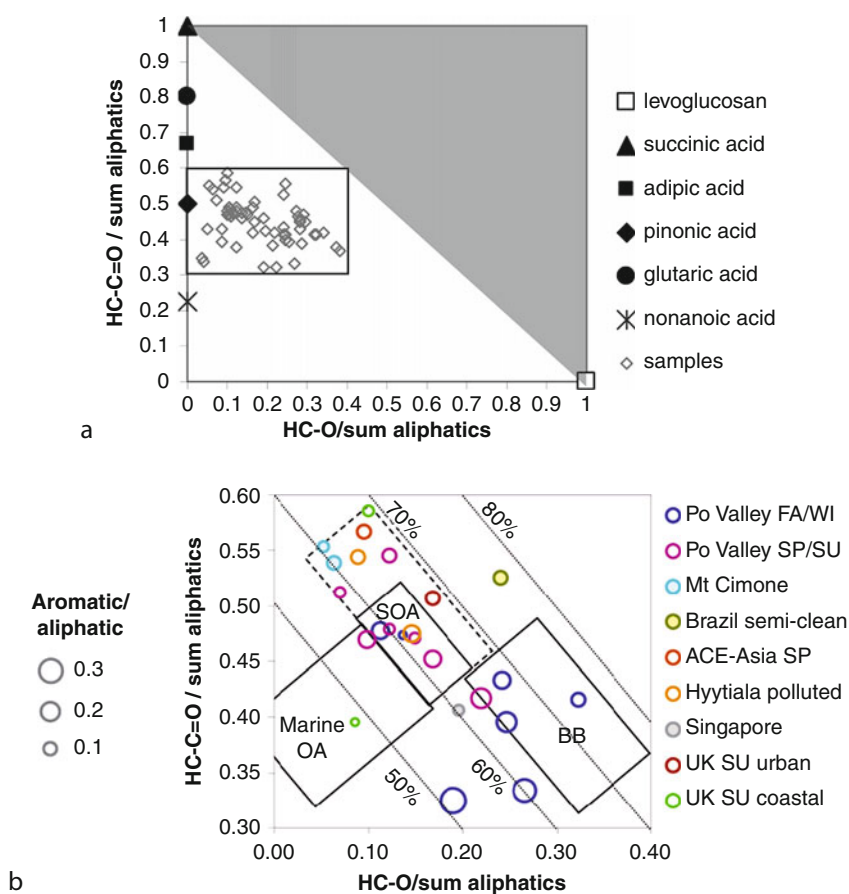
humidities, it is more likely to evaporate. The chemistry of the preexisting aerosol also influences the gas phase partitioning of nitric acid. For example, in the presence of alkaline aerosol species such as sea-salt and dust, nitric acid will preferentially condense to these aerosol types. Since the bulk of sea-salt and dust aerosol are found in supermicron size ranges, it is in these size ranges where nitric acid can also be found. However, in polluted regions, and in the absence of sea-salt and dust, nitric acid is also found in significant amounts in submicron-sized aerosol. Nitric acid and sulfate aerosol are generally sourced from pollution sources; however, in marine regions, the oxidation of dimethylsulfide, a waste product from plankton activity, can provide a significant source of natural  $\text{SO}_2$  and methane sulfonic acid, leading to natural aerosol production.

Aqueous phase reactions can occur in liquid aerosol particles but typically, the majority of aqueous phase production occurs in cloud droplets. It is estimated that between 75% and 90% of aerosol sulfate ( $\text{SO}_4^{--}$ ) is produced through this pathway. In cloud, the droplet surface area and volume of the cloud droplets promote the dissolution of  $\text{SO}_2$  into the aqueous phase where it can be rapidly oxidized by ozone and/or hydrogen peroxide [29]. The oxidation route by ozone is highly pH dependent and dominant in alkaline drops such as those formed on sea-salt aerosol, while the hydrogen peroxide route dominates at lower pH values [30]. When sea-salt is present in cloud nuclei, the cloud behaves as a virtual buffer, increasing alkalinity and accelerating ozone-driven aqueous phase oxidation of  $\text{SO}_2$  to produce sulfate [30]. This sulfate formation route led to the significant production of acidic sulfate aerosol in polluted air and was the primary cause of the acid-rain problem in the 1970s–1980s.

Oxidation of volatile organic compounds (VOCs) leads to the production of condensable organic vapors, ultimately producing secondary organic aerosol (SOA). The major oxidizing species are the OH radical, ozone, and the nitrate radical ( $\text{NO}_3$ ). Like in the case of nitric acid, some of these condensable vapors can be semivolatile. There are both natural VOCs (biogenic: BVOCs) and anthropogenic VOCs (AVOCs), although it is estimated that the BVOC emissions are of the order of ten times that of AVOC emissions. Some modeling

studies suggest that in certain heavily populated regions, AVOCs can contribute significantly to total organic carbon levels [31]; other studies using carbon isotope analysis suggest that BVOC oxidation products contributed to the majority of SOA even in urban regions [32]; however, it should be noted that even BSOA production is somewhat linked to anthropogenic activity as the main VOC oxidants are associated with anthropogenic emissions.

For BVOC emissions, terpenes and sesquiterpenes from terrestrial vegetation emissions, are thought to be the dominant precursors of BSOA, and to a lesser extent isoprene (particularly under high  $\text{NO}_x$  conditions [33]), while aromatic compounds such as toluene, xylene, and trimethyl benzene are thought to account for the majority of ASOA formation. There has been considerable effort to identify the composition of SOA in recent years; however, the vast majority of compounds remain unidentified due to complex mixtures of compounds often with few distinguishing features. SOA is generally oxygenated and water soluble, in contrast to hydrocarbon-like water-insoluble POA. Various approaches have been taken to speciate the organic mass found in SOA. For example,  $^1\text{H}$  nuclear magnetic resonance (HNMR) techniques have been applied to evaluate functional groups in water-soluble OA for different aerosol source regimes [34]. The technique readily detects four main functional groups in atmospheric water-soluble OA: unfunctionalized alkyls ( $\text{H-C}$ ); aliphatic carbons bound to an unsaturated carbon atom ( $\text{H-C-C=}$ ); aliphatic alcohols, ethers, and esters ( $\text{H-C-O}$ ); and aromatic rings ( $\text{H-Ar}$ ). Additionally, formyl ( $\text{H-C=O}$ ), acetal ( $\text{O-CH-O}$ ), and sulfonic ( $\text{H-C-S=O}$ ) groups are detected. Water-soluble OA samples from the boreal forest in Finland (terrestrial biogenic), biomass burning events in Brazil, rural UK, and the Northeast Atlantic (marine) were analyzed for functional groups using this approach. Figure 5 illustrates, on the vertical axis, the aliphatic carbon fraction (carbonylic/carboxylic,  $\text{HC-C=O}$  groups) and, on the horizontal axis, the fraction of aliphatic carbon accounted for by hydroxyl groups for a range of standard SOA compounds (levoglucosan, succinic acid, adipic acid, pinonic acid, glutaric acid, and nonanoic acid) alongside the diverse atmospheric samples. The analysis shows that marine



**Aerosol in Global Atmosphere. Figure 5**

**(a)** Functional group distributions of standards of water-soluble organic compounds. The area including the composition of the real atmospheric samples is also indicated. **(b)** Functional group distribution of WSOCs samples characteristic of specific aerosol sources. Diagonal lines represent the percentage fraction of total oxygenated groups ( $\text{H-C-O} + \text{HC-C=O}$ ) (Copyright American Chemical Society 2007. Reprinted from [34])

and terrestrial biogenics are characterized by relatively high H-C and low H-C-O groups which partly overlap; however, the marine samples have a significantly lower aromatic content. The functional composition of biogenic OA and that of rural anthropogenic OA are relatively close with the rural samples having a higher H-C-O content. Both also contain high fractions of aliphatic carbonyls. The biomass burning water-soluble OA contained the highest content of hydroxyls and the lowest amount of alkyl groups.

The formation of SOA is not always via the gas phase formation of condensable vapors. In the case of isoprene-related SOA production, its relatively high

vapor pressure oxidation products suggest little contribution to SOA; however, alternative routes have been suggested, which produce low-volatility aerosol phase products. For example, tetrols with the same backbone as isoprene have been observed in atmospheric aerosols [35] possibly formed by the heterogeneous acid-catalyzed oxidation of isoprene in the presence of peroxides [36]. Additionally, heterogeneous reaction of isoprene under highly acidic conditions was also observed to lead to the formation of polymeric, humic-like substances [37]. Other studies showed that some high-volatility carbonylic oxidation products can contribute to SOA through the interaction

with highly acidic inorganic preexisting aerosols [38], while polymerization of second-generation products has been suggested to contribute to SOA growth [39]. Cloud modeling studies [40] also predict that low-volatility water-soluble oxidation products of isoprene can be oxidized in the aqueous phase in clouds to produce dicarboxylic acids.

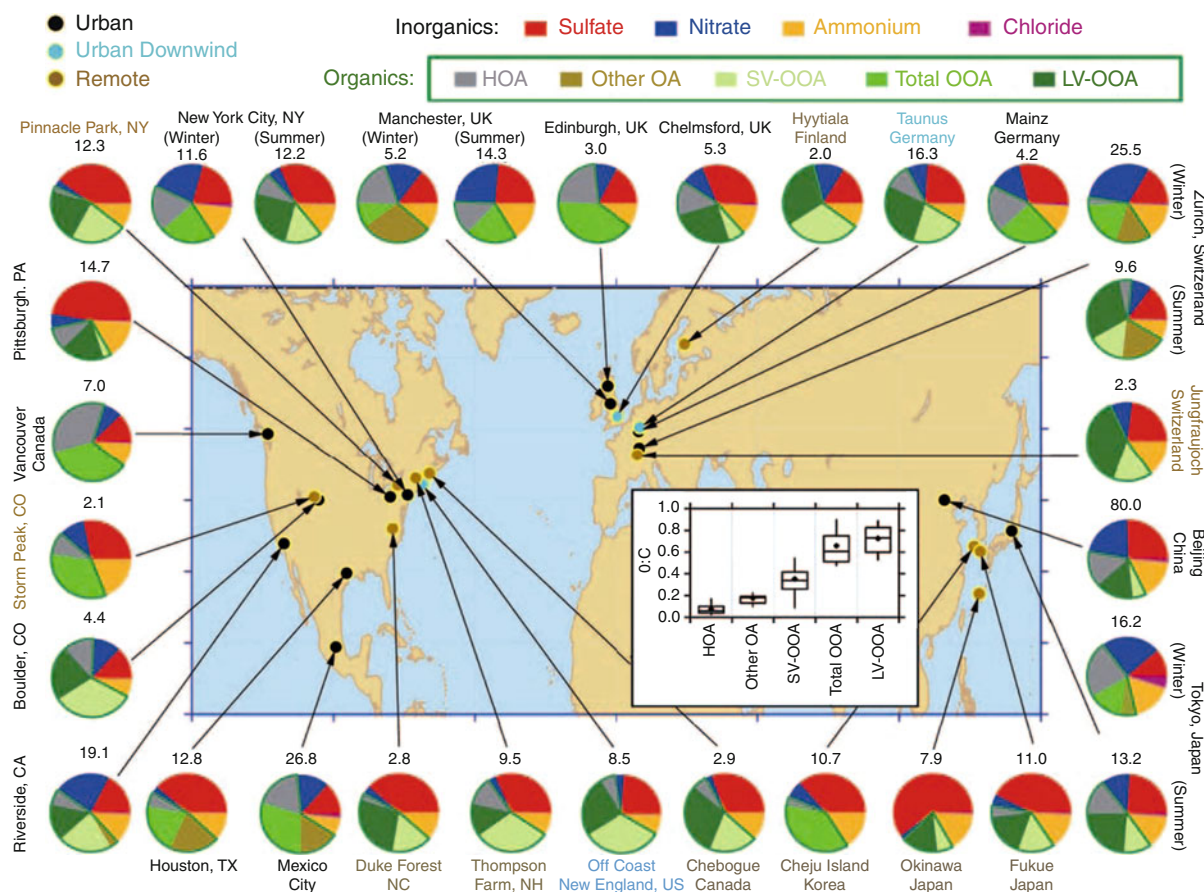
Once SOA is formed, it contributes to a dynamic thermodynamic and chemical evolution system. SOA species can be repartitioned to the gas phase and undergo polymerization and oxidation leading to significant changes in the chemical nature of SOA. The evolution of SOA is well illustrated through the examination of SOA mass spectral properties moving from strong urban source regions to remote continental regions. Aerosol mass spectrometry (particularly using the Aerodyne Aerosol Mass Spectrometer – AMS) delivers increased speciation capability on both POA and SOA. This technology splits the organic aerosol into three generic types – an oxygenated organic aerosol (OOA), a hydrogenated fossil-fuel-derived organic aerosol (HOA), and a biomass burning organic aerosol (BBOA). OOA is characterized by its highly oxidized state with atomic O:C ratios between 0.25 and 1 and can be further categorized into two subgroups based on volatility and O:C ratio. Low-volatility OOA ( $\sim\text{C}_8\text{O}_{5.5}\text{H}_{10}$ ) has a high O:C ratio and is frequently correlated to other secondary species such as sulfate aerosol and is consistent with an aged aerosol. Semivolatile OOA ( $\sim\text{C}_8\text{O}_3\text{H}_{10}$ ) has a lower O:C ratio and is better correlated with other semivolatile species such as nitrates and is consistent with a partially photochemically aged aerosol.

A compilation [41] of measurements of HOA and OOA types, along with major inorganic species, at different urban, urban downwind, and remote rural locations, is shown in Fig. 6. What is clearly evident from the suite of measurements is that organic matter comprises a significant fraction of the aerosol mass at all locations; however, the further from pollution sources, the more oxygenated the organic aerosol is (i.e., the ratio of OOA/HOA increases) and the more aged, or distant from pollution sources, the higher fraction of OOA is low-volatility OOA. This study also reports the evolution of the Mexico City pollution plume within which there is intense SOA formation. They found that near the plume source region,

semivolatile OOA dominated the organic aerosol signal; however, as the plume advected and aged downwind, the fraction of low-volatility OOA and the O:C ratio increased. Laboratory studies on the evolution of biogenic SOA (from  $\alpha$ -pinene oxidation), biomass burning smoke, and the POA surrogate squalane also exhibit similar aging patterns as in the reported atmospheric plumes. The laboratory results demonstrate that in the early stages of SOA production, the OA mass spectrum resembles that associated with the respective source, but as oxidation and aging proceeds, the SOA spectra evolve into initially semivolatile OOA spectra and ultimately into low-volatility OOA spectra. The combined atmospheric and laboratory studies suggest that with atmospheric oxidation, most types of OA converge to a final state of low-volatility OOA, irrespective of the original source. It should, however, be noted that primary and secondary marine aerosol aging was not evaluated and that these, potentially, could behave differently to the terrestrial biogenic and anthropogenic OA types.

The previous study developed, on the basis of the mentioned experimental studies, a simple 2D basis (2D-VBS) model encapsulating the evolution of OA volatility (determined as saturation concentration  $C^*$  at 298K) and oxidation state (O:C ratio) to describe the evolution of OA in the atmosphere [41]. These two properties can be measured in real time and are thus used to constrain the model with experimental field data. The model incorporates photochemical aging using a branching ratio between functionalization and fragmentation kernels and homogeneous and heterogeneous oxidation by OH, and where the first generation of oxidation is explicitly modeled while the second generation is phenomenologically modeled. The conceptual framework was applied to SOA production from  $\alpha$ -pinene ozone-driven oxidation and was found to predict first-generation products with  $1 < C^* < 10^7 \mu\text{g m}^{-3}$  and  $0.1 < \text{O:C} < 0.4$ . Given that generally organics with  $C^* < 10 \mu\text{g m}^{-3}$  will partition into the aerosol phase, it is evident that most of the oxidation products remain as volatile gases. The first-generation products can then functionalize through OH reactions resulting in a tripling of SOA mass production after second-generation oxidation. The model is able to predict the evolution of SOA in the 2D volatility and O:C space, remaining consistent with the production key





**Aerosol in Global Atmosphere. Figure 6**

Total mass concentration (in micrograms per cubic meter) and mass fractions of nonrefractory inorganic species and organic components in submicrometer aerosols measured with the AMS at multiple surface locations in the Northern Hemisphere (21). The organic components were obtained with FA-AMS methods (3, 15–17). In some studies, the FA-AMS methods identified one OOA factor, whereas in other locations, two types, SV-OOA and LV-OOA, were identified. HOA is a surrogate for urban primary OA, and Other OA includes primary OAs other than HOA that have been identified in several studies, including BBOA (Reprinted from [41])

products such as the first-generation cis-pinonic acid, regarded as a semivolatile OOA, and second-generation products such as  $\alpha,\alpha$ -dimethyltricarballic acid, a low-volatility OOA. Similar evolution was found for other SOA sources like evaporated diesel and biomass burning. This 2D volatility and O:C conceptual has made considerable advances in simplifying the treatment of an incredibly complex system and can be further developed and extended to include other key processes such as oligomerization. The combined perspective to emerge from the recent mass spectral measurements and the modeling framework is that most OA appears

to be an intermediate or transition state between primary emissions and highly oxidized volatile gases such as CO and CO<sub>2</sub>.

### Global Aerosol Budgets, Burdens, and Spatial Distributions

The diverse nature of aerosol sources from natural and anthropogenic emissions, different formation and removal processes, and different transport mechanisms results in great inhomogeneity in annual emission budgets, atmospheric burdens, and spatial distributions of individual aerosol species. The result is ultimately

significant variability in radiative forcing and air pollution on regional scales. To quantify budgets and distributions, general circulation (climate) models or chemical transport models, are used in conjunction with emission inventories. These models contain aerosol–cloud chemistry formation and evolution modules along with deposition schemes and are underpinned by field and laboratory studies into key processes. In addition, quantification is achieved by long-term observations of atmospheric composition at various point locations around the globe. Remote sensing of aerosol parameters from surface and space-borne platforms also provide added information on horizontal and vertical distributions.

### Global Budgets

Atmospheric aerosols are derived from both natural and anthropogenic sources. Natural sources include sulfate and sea spray (sea-salt and primary organic matter) from the oceans, dust from desert regions, particulate organic matter from biogenic sources (both terrestrial and oceanic), and sulfate and silicates from volcanoes. Anthropogenic sources include sulfate, black carbon and primary and secondary organic matter from fossil fuel combustion and biomass burning, dust from industrial sources, and nitrates mainly from transport emissions. The best estimated emissions and resulting atmospheric mass and number burdens are displayed in Table 1 [42]. The largest natural emission is that of sea-salt with a “best guess” of  $10,130 \text{ Tg a}^{-1}$ , followed by dust with a best guess emission of  $1,600 \text{ Tg a}^{-1}$ . Biogenic secondary organic emissions are  $25 \text{ Tg a}^{-1}$ , compared to  $3.5 \text{ Tg a}^{-1}$  for anthropogenic emissions. For sulfate aerosol, anthropogenic emissions ( $122 \text{ Tg a}^{-1}$ ) dominate over natural biogenic ( $57 \text{ Tg a}^{-1}$ ) and volcanic ( $21 \text{ Tg a}^{-1}$ ) emissions and for primary organic aerosol, anthropogenic emissions amount to  $95 \text{ Tg a}^{-1}$  while natural oceanic primary organic aerosol (not shown in Table 1) is estimated to be  $16 \text{ Tg a}^{-1}$  [56]. Black carbon emissions from fossil fuel amount to  $4.5 \text{ Tg a}^{-1}$ .

### Global Burdens

The spatial distribution of total aerosol mass is shown in Fig. 7. The distribution is a medium mass composite

from 10 global atmospheric climate (general circulation) models and chemical transport models for the year 2000. From the model output, the global average column burden for all aerosol types is  $\sim 50 \text{ mg m}^{-2}$ , with regional loadings exceeding  $500 \text{ mg m}^{-2}$ , in places such as North Africa and Asia. These plumes of high aerosol burdens extend out into ocean regions (e.g., the African plume is seen to extend out west into the Atlantic). North America and Europe exhibit moderate levels of aerosol burden, while somewhat higher burdens are seen over South America and Australia. Significant burdens are also seen over remote oceanic regions such as the south Pacific and Indian Ocean. The dominant source regions for different species are revealed when burden distribution maps are examined for individual aerosol species. For black carbon, the global burden is  $0.36 \text{ mg m}^{-2}$ , with values exceeding  $5 \text{ mg m}^{-2}$  occurring over China and burdens of  $1\text{--}2 \text{ mg m}^{-2}$  occurring over regions such as Africa and South America, with slightly lower burdens occurring over Europe and North America. Significant black carbon plumes are seen to advect out over the Atlantic and Indian Oceans from the two aforementioned highest burden regions.

For dust, the global average burden is  $\sim 25 \text{ mg m}^{-2}$ , with peak burdens exceeding  $500 \text{ mg m}^{-2}$  occurring over the Sahara desert in Northwest Africa. East Europe and Asia are also seen as significant dust source regions, and to a lesser degree, western USA and central Australia. For Particulate Organic Matter, the spatial distribution of the column burden is similar to that of black carbon, but with a significantly higher column burden of  $3 \text{ mg m}^{-2}$ . Peak values approach  $50 \text{ mg m}^{-2}$  and the main source regions are the southwestern African Continent, South America, Europe, India and China, northern Australia, and southeast USA. Significant levels of the order of  $1 \text{ mg m}^{-2}$  are also seen over most oceanic regions. Non-sea-salt sulfate exhibits a global average burden of  $\sim 3.5 \text{ mg m}^{-2}$ , with peak values of the order of  $50 \text{ mg m}^{-2}$ . Again, the main developed and developing countries with significant population densities appear as burden “hot spots” as for the other anthropogenic aerosol types (e.g., black carbon and particulate organic matter). In contrast to other anthropogenic sources, there are notable burdens over remote oceanic regions, reflecting also its natural oceanic source.

**Aerosol in Global Atmosphere. Table 1** Particle emission/production and burdens estimated for the year 2000  
(Copyright Elsevier 2008. Reprinted from [42])

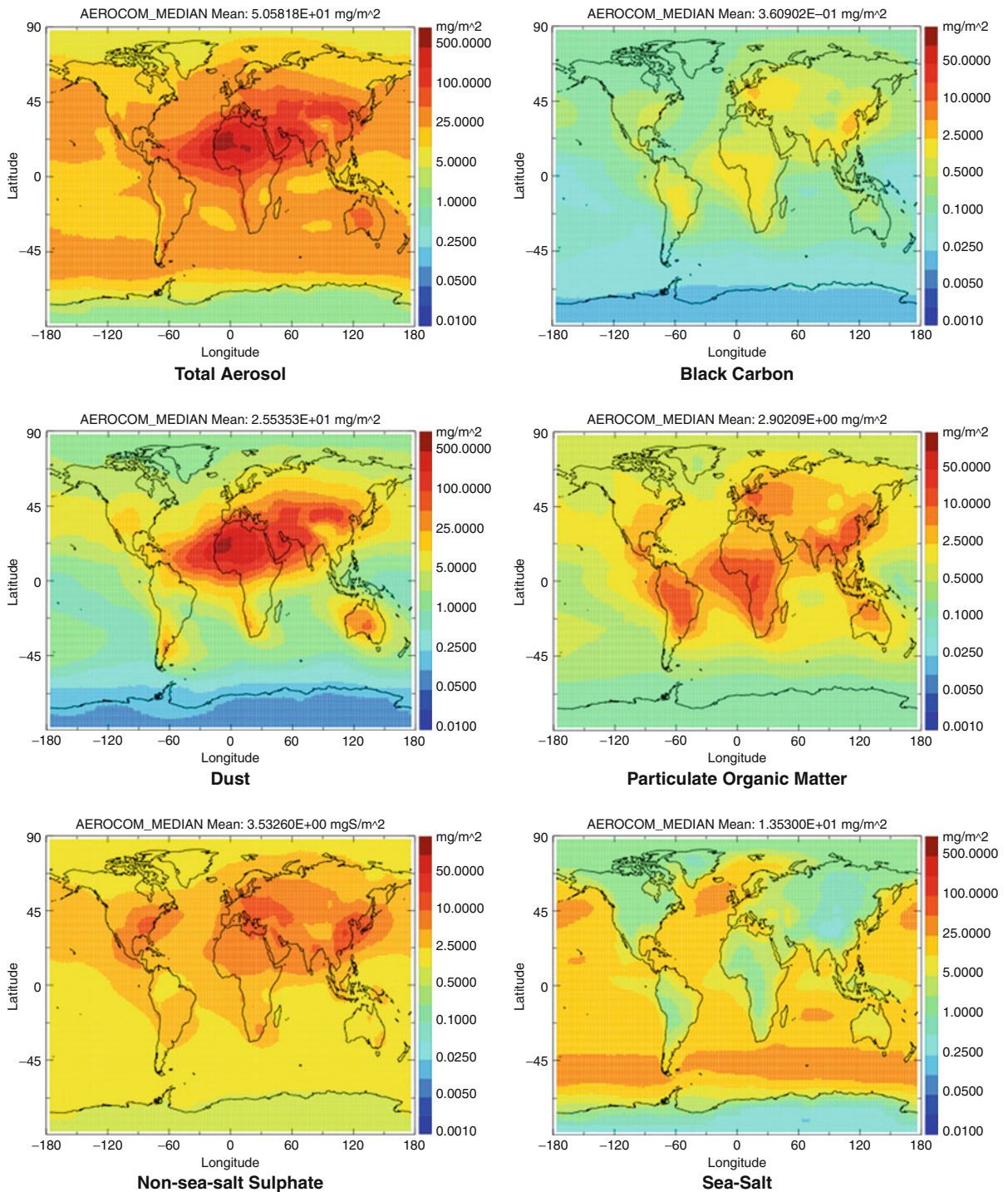
A

|                          | Mass emission |                    |        | Mass burden | Number prod.         | Number burden        |
|--------------------------|---------------|--------------------|--------|-------------|----------------------|----------------------|
|                          | “Best guess”  | Min                | Max    | Tg          | a <sup>-1</sup>      |                      |
|                          |               | Tg a <sup>-1</sup> |        |             |                      |                      |
| Carbonaceous aerosols    |               |                    |        |             |                      |                      |
| Primary organic (0–2 μm) | 95            | 40                 | 150    | 1.2         | –                    | 310·10 <sup>24</sup> |
| Biomass burning          | 54            | 26                 | 70     | –           | 7·10 <sup>27</sup>   | –                    |
| Fossil fuel              | 4             | 3                  | 9      | –           | –                    | –                    |
| Biogenic                 | 35            | 15                 | 70     | 0.2         | –                    | –                    |
| Black carbon (0–2 μm)    | 10            | 8                  | 14     | 0.1         | –                    | 270·10 <sup>24</sup> |
| Open burning and biofuel | 6             | 5                  | 7      | –           | –                    | –                    |
| Fossil fuel              | 4.5           | 3                  | 6      | –           | –                    | –                    |
| Secondary organic        | 28            | 2.5                | 83     | 0.8         | –                    | –                    |
| Biogenic                 | 25            | 2.5                | 79     | 0.7         | –                    | –                    |
| Anthropogenic            | 3.5           | 0.05               | 4.0    | 0.08        | –                    | –                    |
| Sulfates                 | 200           | 107                | 374    | 2.8         | 2·10 <sup>28</sup>   | –                    |
| Biogenic                 | 57            | 28                 | 118    | 1.2         | –                    | –                    |
| Volcanic                 | 21            | 9                  | 48     | 0.2         | –                    | –                    |
| Anthropogenic            | 122           | 69                 | 214    | 1.4         | –                    | –                    |
| Nitrates                 | 18            | 12                 | 27     | 0.49        | –                    | –                    |
| Industrial dust etc.     | 100           | 40                 | 130    | 1.1         | –                    | –                    |
| Sea-salt                 |               |                    |        |             |                      |                      |
| d < 1 μm                 | 180           | 60                 | 500    | 3.5         | 7.4·10 <sup>26</sup> | –                    |
| d = 1–16 μm              | 9,940         | 3,000              | 20,000 | 12          | 4.6·10 <sup>26</sup> | –                    |
| Total                    | 10,130        | 3,000              | 20,000 | 15          | 1.2·10 <sup>27</sup> | 27·10 <sup>24</sup>  |
| Mineral (soil) dust      |               |                    |        |             |                      |                      |
| <1 μm                    | 165           | –                  | –      | 4.7         | 4.1·10 <sup>25</sup> | –                    |
| 1–2.5 μm                 | 496           | –                  | –      | 12.5        | 9.6·10 <sup>25</sup> | –                    |
| 2.5–10 μm                | 992           | –                  | –      | 6           | –                    | –                    |
| Total                    | 1,600         | 1,000              | 2,150  | 18 ± 5      | 1.4·10 <sup>26</sup> | 11·10 <sup>24</sup>  |

Range reflects estimates reported in the literature. The actual range of uncertainty may encompass values both larger and smaller than those reported here. Values are based on the following publications: [31, 43–55])

Finally, the sea-salt burden is 13.5 mg m<sup>-2</sup>, with peak burdens of ~50 mg m<sup>-2</sup>. The peak burdens occur at midlatitudes, or the “roaring 40s,” north and south of the equator, and are associated with regions of intensive cyclonic activity.

It is interesting to note that in Table 1, sea-salt was highlighted as being the single most important aerosol emission type, but it had a mass burden slightly lower than dust which had significantly lower emissions than sea-salt. In the AEROCOM (<http://dataipsi.ipsl.jussieu>).



**Aerosol in Global Atmosphere. Figure 7**

Median global distribution of aerosol column mass, total and per aerosol species, derived from 10 AEROCOM models for the year 2000 (Copyright European Geophysical Union 2006. Reprinted from [57])



fr/AEROCOM; [57]) study, dust also has a higher burden for generally lower emissions. The explanation is that significant dust plumes are vented into the Free Troposphere where removal rates are slower and residence times are longer than in the marine boundary layer where removal rates are likely higher due to higher precipitation rates and, for the larger-sized particles, shorter gravitational settling timescales (i.e., sea-salt is not injected at as high altitudes as dust, so they settle out more rapidly).

At any one location, the chemical composition of atmospheric can vary significantly depending on the air mass back trajectory arriving at that location. An example of the chemical variability encountered is demonstrated for the Northeast Atlantic region where clean marine and polluted continental air masses are frequently encountered at the Mace Head Atmospheric Research Station, which is representative of the region. Figure 8 illustrates the absolute chemical composition and the relative chemical composition observed in two contrasting air masses (continental polar, maritime polar) of different origin over the Northeast Atlantic. Typically, the main ionic compounds such as sodium, chloride, nitrate, sulfate, and ammonium are easily detected; however, it is also possible to quantify the water-soluble and water-insoluble organic component. It is readily seen that the absolute and relative contributions of the above species varies greatly with air mass, or source origin [58]. Peak submicron mass loadings occur for the continental air, while peak supermicron masses occur for marine air. In marine polar air, the supmicron sizes are almost exclusively composed of sea-salt, while submicron sizes are marginally dominated by nss-sulfate, with the second largest contribution derived from sea-salt. Small amounts of ammonia, water-soluble organic matter and water-insoluble organic matter are also evident. By contrast, in the continental air, sea-salt and nitrate are the two dominant species in supermicron sizes, with smaller amounts of organic matter visible and in the submicron sizes, nss-sulfate and water-soluble organic matter dominate. It should be noted that marine aerosol can be significantly enriched in organic matter depending on the degree to which air masses pass over biologically rich waters.

Mass is only one important parameter of an aerosol size distribution. Perhaps the other most important

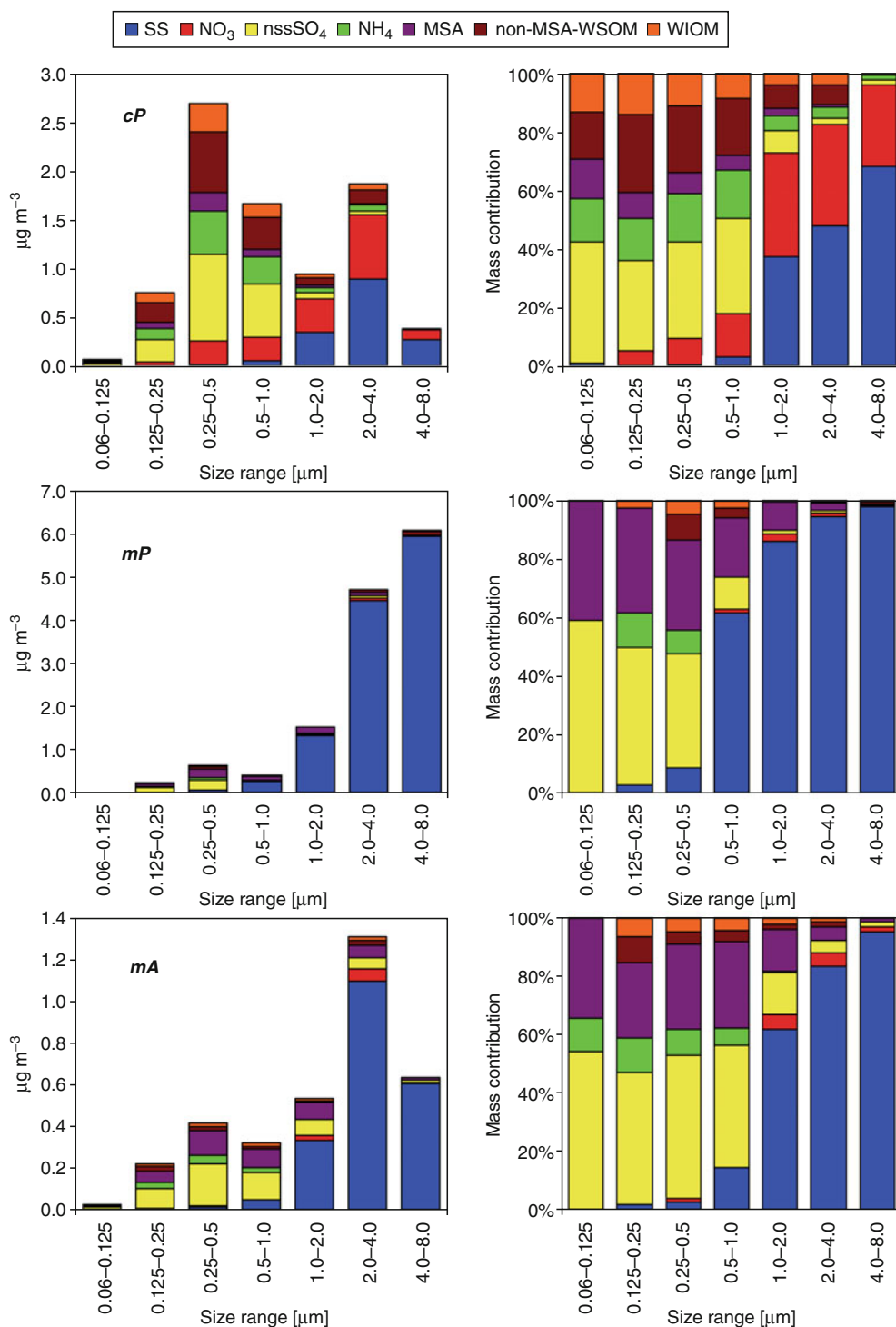
parameter is the number concentration, and in particular, the concentration in the so-called accumulation mode (100–500 nm diameter). This size range is not only important for scattering and radiative effects of atmospheric aerosols, it is crucially important in terms of influencing the number concentration of cloud condensation nuclei. Figure 9 illustrates the horizontal and vertical (zonal) average concentration in the accumulation mode size range [59]. It can be seen that the highest number concentration of aerosols ( $N = 3,000\text{--}5,000\text{ cm}^{-3}$ ) occurs over regions in Europe such as the Netherlands and Northern Italy, China, India, and Brazil.

### Vertical Distributions

The vertical distribution of aerosol particles, measured as aerosol backscatter using a LiDAR (Light Detection And Ranging), typically reveals maximum particle mass and number concentrations in the atmospheric boundary layer (lower  $\sim 1,000\text{--}2,000\text{ m}$ ) as shown in Fig. 10. In this figure, the attenuated atmospheric aerosol backscatter profiles illustrate the vertical aerosol structure [58]. In the case shown, the main boundary layer temperature inversion is at  $\sim 1,500\text{ m}$  and within the boundary layer there is a surface-mixed layer at  $500\text{ m}$ . The peak aerosol burden occurs in the surface-mixed layer and corresponds to a European pollution plume advecting out into the Atlantic. Significant aerosol backscatter (and hence aerosol loading) is seen in the residual boundary layer between  $1,000$  and  $2,000\text{ m}$ . Above  $2,000\text{ m}$ , in the free troposphere, very clean aerosol levels are seen.

Figure 11 displays a 10-year climatology of accumulation mode zonal vertical distributions demonstrating that the Northern Hemisphere from  $60^\circ\text{N}$  to  $20^\circ$  into the Southern Hemisphere contains the highest particle number concentrations ( $\sim 700\text{ cm}^{-3}$ ). The cleanest concentrations occur in and around Antarctica, while the second cleanest is in the upper troposphere–stratosphere region.

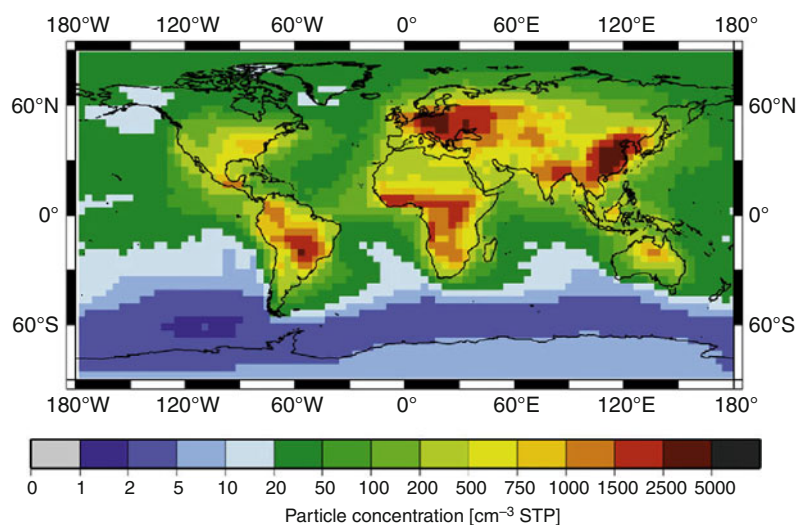
Vertical profiles of total number of concentrations (historically defined as condensation nuclei, or CN, typically larger than  $10\text{ nm}$ ), and size-resolved concentration measurements (with radius  $> 0.15\text{ }\mu\text{m}$  to sizes greater than  $2\text{ }\mu\text{m}$ ) up to the higher stratosphere are shown in Fig. 12 (taken from [60]). In this figure,



**Aerosol in Global Atmosphere. Figure 8**

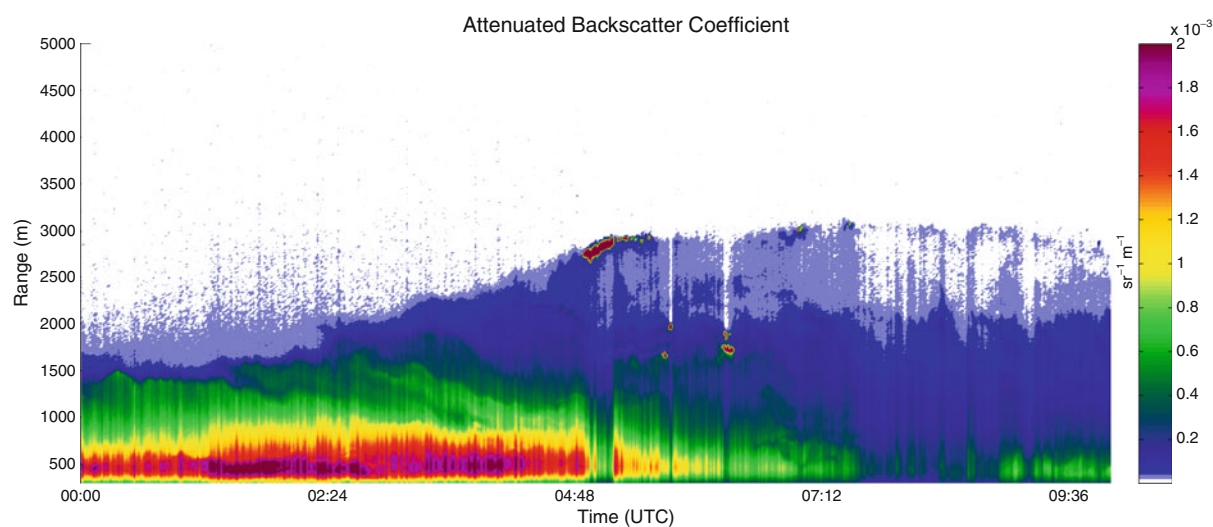
Absolute (left panel) and relative (right panel) chemical composition of the aerosol samples collected in continental polar (cP), maritime polar (mP), and maritime Arctic (mA) air masses measured at Mace Head, Northeast Atlantic, during June 2008 (Copyright European Geophysical Union 2009. Reprinted from [58])





**Aerosol in Global Atmosphere. Figure 9**

Climatological annual means of simulated accumulation mode aerosol particle concentration obtained from a 10-year integration. The geographical distribution is of the lowermost model layer, which is about 60 m thick (Copyright European Geophysical Union 2006. Reprinted from [59])

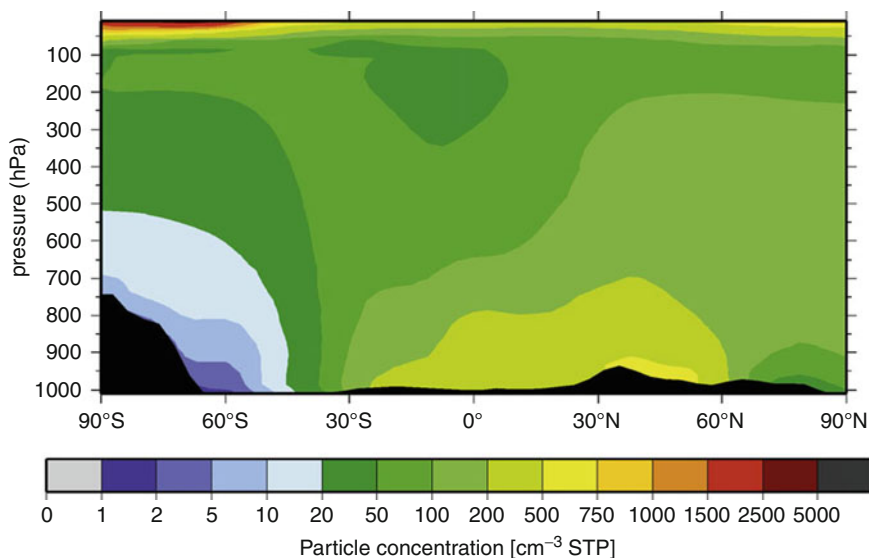


**Aerosol in Global Atmosphere. Figure 10**

LIDAR aerosol attenuated backscatter vertical profiles of the atmospheric boundary layer and free troposphere. Measurements were taken at the Mace Head Atmospheric Research Station on the Irish Atlantic coastline. Measurements were taken during polluted continental air outflow from Europe into the Atlantic region

concentrations are shown for 1 year and 15 years after the Mt. Pinatubo eruption. What is evident is that total aerosol, or CN, concentration at the surface is between  $500 \text{ cm}^{-3}$  (typical of clean background continental air masses) and  $3,000 \text{ cm}^{-3}$  (typical of polluted

continental air masses) and that free troposphere, or upper free tropospheric concentrations are relatively constant at a concentration of the order of  $500 \text{ cm}^{-3}$ . Above 15 km, in the stratospheric layer, CN concentrations are typically  $1\text{--}10 \text{ cm}^{-3}$ , despite the volcanic



**Aerosol in Global Atmosphere. Figure 11**

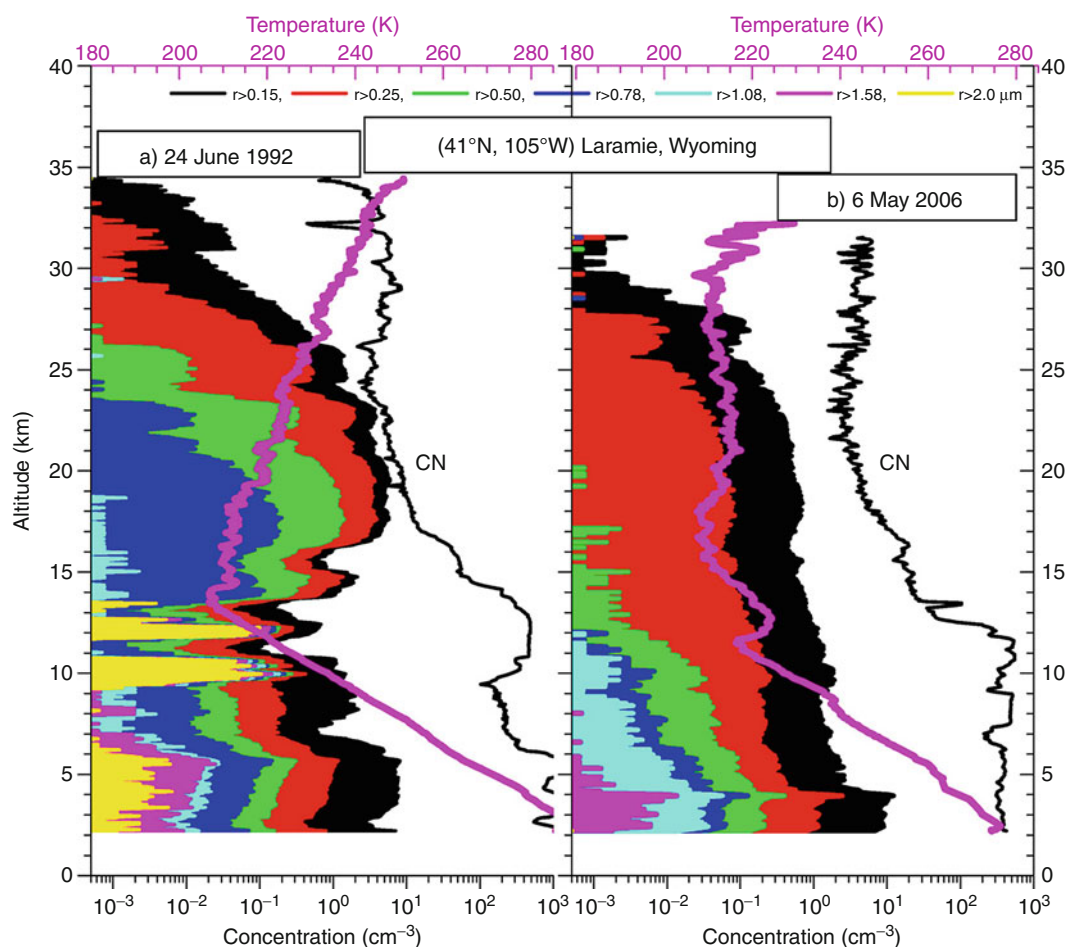
Climatological annual means of simulated zonally averaged vertical cross-section accumulation mode aerosol particle concentration obtained from a 10-year integration. The concentrations are given at STP conditions (Copyright European Geophysical Union 2006. Reprinted from [59])

eruption. However, what is evident in the difference in time lapse following the eruption is that there is a significant increase in aerosol concentration (at least an order of magnitude) at sizes greater than  $0.15\ \mu\text{m}$  radius 1 year after the eruption compared to what is considered the background stratospheric aerosol burden 15 years after the eruption. Clearly, volcanic eruptions are a major source of stratospheric aerosols with injections to the tune of  $30\ \text{Tg SO}_2$  for intense eruptions such as Pinatubo, producing the main precursor for stratospheric aerosol formation. In terms of stratospheric aerosol chemical composition, the majority of stratospheric aerosol is composed of 75% sulfuric acid and 25% water, formed from binary homogeneous nucleation and subsequent condensation of sulfuric acid as suggested by boiling point measurements of the stratospheric aerosol [61]. Other sources of stratospheric aerosols include organic carbonyl sulfide, particularly above altitudes of 25 km, and meteoritic material, rocket exhaust, and aircraft emissions are regarded as having minimal influence on the stratospheric aerosol layer [[60] and references therein.].

A particularly interesting stratospheric aerosol is aerosol forming polar stratospheric clouds (PSCs). These aerosol haze layers, or more often known as

cloud layers, have been the subject of intense interest after the discovery of their role in stratospheric ozone loss through converting chlorine inactive to active chlorine and denitrifying the stratosphere. Observational, laboratory, and theoretical work has since shown that PSC particles above the ice point are either solid nitric acid trihydrate (NAT) particles or liquid ternary aerosol (LTA), droplets composed of nitric acid, sulfuric acid, and water. The composition and phase of the particles determines the threshold temperatures at which they occur and persist – specifically, NAT will exist at temperatures 3–5 K above LTA temperatures, while LTA will begin to appear about 3 K above the ice point for polar stratospheric levels of gaseous nitric acid and water [62, 63]. Particle phase also controls the development of particle size, which plays a large role in the extent and rate of dehydration/denitrification, thus determining the time frame during which chlorine will remain active, and ultimately determining the fundamental processes associated with ozone depletion in the polar stratosphere.

The main open question concerning the formation of PSCs is the process by which NAT forms. LTA growth is strictly a function of gas-phase-mixing ratios and temperature, as the vapor pressures of sulfuric acid,



**Aerosol in Global Atmosphere. Figure 12**

Stratospheric aerosol profiles above Laramie Wyoming ( $41^{\circ}\text{N}, 106^{\circ}\text{W}$ ) indicating the concentration of condensation nuclei (CN) and aerosol with radii greater than 0.15, 0.25, 0.50, 0.78, 1.08, 1.58, and  $2.0\ \mu\text{m}$ . Temperature with scale at the top is also shown. These profiles were measured in (a) June 1992, 1 year after the Pinatubo eruption and (b) May 2006, 15 years after Pinatubo (Copyright Elsevier 2008. Reprinted from [60])

nitric acid, and water adjust to environmental temperature and molecular collisions. LTA begins to grow on the SA as soon as temperatures reach about 3 K above the ice point. At these temperatures, the concentration of NAT particles is often at best only 0.001 of the aerosol population [64–66] indicating clearly a nucleation barrier which is not well understood. Laboratory investigations suggest that LTA must be significantly supercooled for NAT nucleation. Arctic PSC observations in the early 2000s suggest the importance of a small population of NAT particles and that such a population forms at temperatures just a few degrees below the NAT point in both leewaves [64]

and synoptic situations [66]. These latter observations are the clearest evidence yet indicating that NAT forms readily on a small fraction ( $10^{-4}$ – $10^{-5}$ ) of SA through heterogeneous nucleation.

### Water Uptake by Aerosols

One of the most important gases in the atmosphere is water vapor as it not only contributes predominantly to the natural greenhouse effect, resulting in a livable global average temperature, in its condensed liquid and ice phases, it forms aerosol haze and cloud layers which contribute to approximately 50% of the planetary albedo. Many aerosol types have an affinity to

water and, under the right humidity conditions, they can become solution droplets, larger than the dry aerosol size, by uptake of water vapor from the air around it. Aerosol humidification has implications for optical scattering, condensation processes, and chemical interactions on and within aerosol particles. Liquid water clouds are extensive in the atmosphere, but as was shown above, pure water droplets cannot form under atmospheric conditions. Condensation nuclei are required for the formation of cloud droplets. Cloud droplets are aerosol particles that have rapidly grown to 10–100 times the size of the initial nucleus, corresponding to more than 1,000 times increase in mass, almost solely through water uptake.

### Aerosol Hygroscopicity

A large fraction of the atmospheric aerosol population is water soluble or partially water soluble. The soluble aerosols are typically salts such as sea-salt or ammonium sulfate which readily dissociates in water to form an electrolyte solution droplet. In subsaturated air (relative humidity < 100%), these solution droplets can exist due to the lowering of the equilibrium vapor pressure above the droplet surface by the presence of the dissolved ions (known as the Raoult Effect). As a result, soluble aerosol particles, depending on the chemical composition, readily form solution drops with a water fraction dependent on relative humidity. This means that the “wet” aerosol equilibrium size responds to, and is dependent on, ambient relative humidity. The “wet” aerosol size will grow with increasing humidity and shrink with reducing humidity. Since the optical properties (e.g., scattering) of aerosols is size dependent (and proportional to cross-sectional area), relative humidity can significantly influence the hazing, or optical extinction effect of aerosol layers. A measure of this effect is the hygroscopic growth factor, normally defined as the change in aerosol diameter from a low relative humidity (e.g., 40%) corresponding to a dehydrated dry-size diameter to a high relative humidity (normally 90%) with significant hydration. Salts like ammonium sulfate and sea-salt have a growth factor of 1.8 and 2.2, respectively, over this humidity range, while organic aerosol possesses a growth factor typically from 1 to 1.4 depending on the degree of solubility.

### Cloud Droplet Activation

In supersaturated environment, a subset of the aerosol can grow to a critical size which becomes metastable, and with an infinitesimal further increase in humidity, the droplet will become unstable and spontaneously grow to sizes of the order of 10  $\mu\text{m}$  and become a cloud droplet. This process is called cloud droplet activation. The description of the equilibrium size of a droplet with water saturation ratio, founded on the early work of Köhler [67], is now well established and can be readily derived from the *Clausius–Clapeyron* equation modified to give a general equilibrium relation between an aqueous salt solution droplet and water vapor (known as the Köhler curve):

$$S = a_w \exp\left(\frac{2v_w\sigma}{RT_r}\right)$$

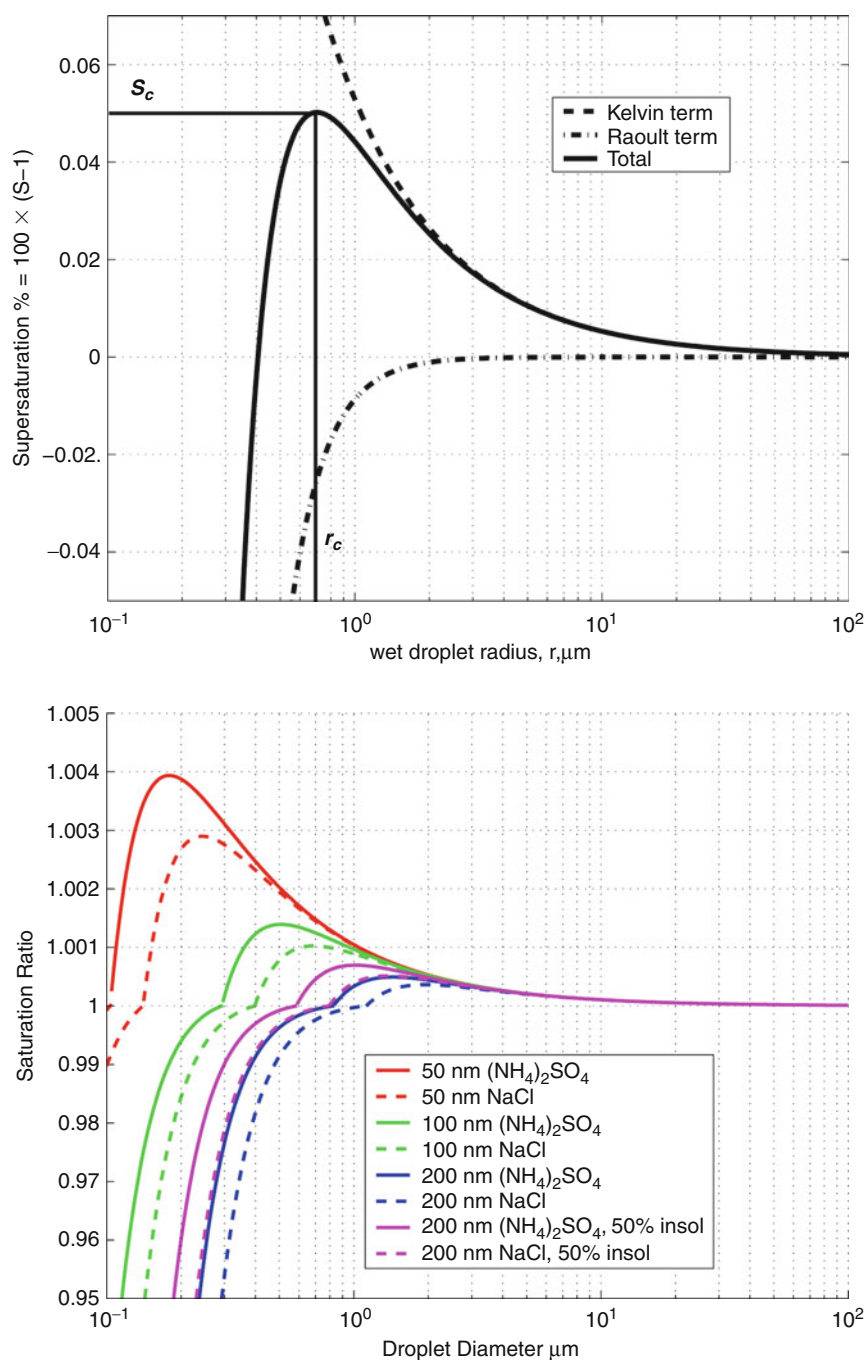
$S$ , is known as the saturation ratio,  $a_w$  is the water activity, the term within the brackets is the Kelvin factor,  $v_w$  is the partial molar volume of water,  $\sigma$  is the surface tension of the solution at the composition of the droplet,  $R$  is the universal gas constant,  $T$  is the droplet temperature, and  $r$  is the particle radius.

This form of the Köhler equation can be simplified to [1, 68–72]

$$S = 1 + \frac{A}{r_c} - \frac{B}{r_c^3} \quad A = \frac{2\sigma}{n_w kT} \quad B = \frac{vmM_w}{M_s^4/3\pi\rho_w}$$

where  $v$  is the number of dissociated ions per solute molecule,  $m_s$  is the solute mass, and subscripts “s” and “w” relate to solute and water properties, respectively. The term in  $A$  is denoted the Kelvin or curvature term, and the term in  $B$ , is known as the Raoult or solute term. The curvature term results in an increase in equilibrium vapor pressure with reducing particle size, while the solute term results in a reduction in equilibrium vapor pressure with increasing solute mass.

Figure 13 shows the contribution of the Kelvin and Raoult terms to equilibrium vapor pressure and equilibrium droplet size for a 200 nm dry diameter ammonium sulfate particle in a subsaturated and supersaturated regime [73]. This form of the expression shows a single characteristic maximum in supersaturation for a given dry composition and size, known as the critical supersaturation,  $S_c$ , associated with



**Aerosol in Global Atmosphere. Figure 13**

*Top:* The Köhler equation can be envisaged as the competition between the curvature (Kelvin) and solute (Raoult) terms.

*Bottom:* Activation curves for a range of dry diameter of salt ( $(\text{NH}_4)_2\text{SO}_4$  – solid, NaCl – dashed) particles (red, green and blue curves) and for 200 nm particles containing 50% by mass insoluble core (magenta) (Copyright European Geophysical Union 2006. Reprinted from [73])



a unique size, denoted by the critical radius,  $r_c$ . Using the simplified Köhler curve expression, the analytical solutions for the critical quantities are:

$$r_c = \sqrt{\frac{3B}{A}} \quad S_c = \left(\frac{4A^3}{27B}\right)^{\frac{1}{2}}$$

For an increasing environmental value of  $S$  below  $S_c$ , there is a unique equilibrium droplet size. Once the droplet grows beyond its critical size (i.e., as the environmental  $S$  increases above  $S_c$ ), the droplet will exhibit unimpeded growth unless the environmental  $S$  reduces below the equilibrium value of  $S_c$  at the instantaneous value of  $r_c$ . In this case, with no further change in  $S$ , the droplet will evaporate to its subcritical equilibrium size.

The Köhler expression can be envisaged as the competition between the two expressions of component properties determining activation of particles: the curvature term and the solute term. The solute terms depend first on the number of solution molecules and then on the dissociation of these molecules. The effect can be illustrated for two frequently assumed cloud condensation nuclei types: ammonium sulfate and sodium chloride.  $(\text{NH}_4)_2\text{SO}_4$  has a molecular weight of 132 g Mol<sup>-1</sup>, while that of NaCl is 58.5 g Mol<sup>-1</sup>. Thus, in the absence of dissociation, a given mass of NaCl in solution would yield 2.26 times more dissolved molecules than  $(\text{NH}_4)_2\text{SO}_4$ . Assuming full dissociation (infinite dilution),  $(\text{NH}_4)_2\text{SO}_4$  yields three ions while NaCl yields two, so the net effect of the molecular mass and dissociation is that NaCl is  $2.26/1.5 = 1.5$  times more active than  $(\text{NH}_4)_2\text{SO}_4$  for the same dry mass of particle (the  $S_c$  ratio for the two species is around 1.22 due to the square root dependence). This is illustrated in Fig. 13 where the peak supersaturation is plotted versus dry diameter for particles comprising each electrolyte. This figure directly illustrates the significant differences in the critical supersaturation as a function of both the chemical composition and dry size of a particle (Raoult and Kelvin effects). Additionally, in this figure is illustrated the effect of nuclei containing an insoluble core. The effect of an insoluble core is to increase the critical supersaturation required for activation. Cloud nuclei can comprise more complex composition compared to the above cases which will require modifications to the Köhler expression. For example, nuclei can contain limited solubility

components, immiscible liquid phases, and surface-active components. In the latter case, nuclei containing surface active components which alter the droplet surface tension (i.e., reduce surface tension) will have a reduced contribution from the Kelvin term in the Köhler equation and consequently will require a lower critical supersaturation for activation. However, critical supersaturation is not always decreased since the surfactant may not dissociate into as many ions and while the Kelvin term may be reduced, so may the Raoult effect to an even greater amount. This effect for an ammonium sulfate nuclei (which dissociates into three ions) enriched with 30% mass fraction by the surfactant stearic acid (single ion) is that the critical supersaturation is actually increased. Supersaturations in natural clouds are of the order of 0.1–0.5% for stratiform clouds and below 1% for most convective clouds. This provides a lower limit to the size of particles that can act as cloud nuclei, namely, 30–50 nm, but more commonly, aerosols larger than 100 nm provide the majority of cloud nuclei. Nitric acid can also influence cloud nucleus activation in that it can temporarily partition into the aerosol phase as a nucleus takes up water vapor in the rising humidity field below cloud base. The uptake of nitric acid into the solution droplet increases the solute mass and, as a result, lowers the critical supersaturation required for the nucleus.

Ice nuclei are also important cloud nuclei but produce cloud particles in a very different manner and are comprised of very different chemical composition. Typically, they comprise non-water-soluble matter such as mineral dust and certain primary biogenic aerosol particles. Ice particles are formed either by heterogeneous nucleation in a supersaturated (with respect to ice) environment or homogeneous freezing of supercooled liquid droplets [74]. Far less is known about the nature of ice nuclei than what is known about cloud condensation nuclei.

### Aerosols, Global Radiative Budget, Precipitation, and Climate

Aerosols affect the global radiative budget, and hence the global climate, directly through the scattering and/or absorption of incoming solar radiation (and for some aerosol species generally of supermicron size, outgoing infrared radiation) and indirectly



through the modification of cloud microphysics and radiative properties. The former effect is called the direct aerosol radiative forcing effect and the latter, the indirect radiative forcing effect where radiative forcing is defined as “The radiative forcing of the surface-troposphere system due to the perturbation in or the introduction of an agent (say, a change in greenhouse gas concentrations) is the change in net (down minus up) irradiance (solar plus longwave in  $\text{Wm}^{-2}$ ) at the tropopause after allowing for stratospheric temperatures to readjust to radiative equilibrium, but with surface and tropospheric temperatures and state held fixed at the unperturbed values.” In simple terms, it is “the rate of energy change per unit area of the globe as measured at the top of the atmosphere” and often it refers to the change since preindustrial conditions (year 1750). Radiative forcing affects the climate, while alterations in aerosol availability and radiative forcing can affect precipitation.

### Direct Effect

Rayleigh scattering describes the irradiance scattered by a sphere as small compared to the incident wavelength and proportional to  $1/\lambda^4$ , while Rayleigh absorption is proportional to  $1/\lambda^4$  [75]. When particles are comparable to the incident wavelength, Mie scattering must be invoked where scattering is strongly dependent on particle size, shape, composition (via refractive index), and orientation relative to the incident radiation. For particles large relative to the wavelength, the extinction efficiency converges at 2 (the extinction paradox where twice the energy is removed compared to that derived using geometrical optics). For particles with a complex refractive index, some fraction of the incident radiation is absorbed. The extinction efficiency is defined as the sum of the scattering and absorption efficiency and the single scattering albedo ( $\omega$ ) is the ratio of the scattering coefficient to the extinction coefficient. One parameter used to quantify scattering extinction is the mass scattering and mass absorption efficiency, which quantifies scattering and extinction in terms of the aerosol mass distribution function. Both the mass scattering and mass absorption efficiencies as a function of particle size exhibit peak scattering and absorption efficiencies at sizes between 0.1 and 1  $\mu\text{m}$  for typical atmospheric aerosol chemical compositions pointing to this size

range as being the most important size range for aerosol optical properties. Efficient aerosol scattering types are sulfates and sea-salt, while efficient absorption types are black carbon and dust constituents such as hematite.

One of the more important parameters associated with the direct radiative effect is aerosol optical depth (AOD) which describes the extinction of radiation as it propagates through the atmosphere:

$$I/I_0 = e^{-\tau}$$

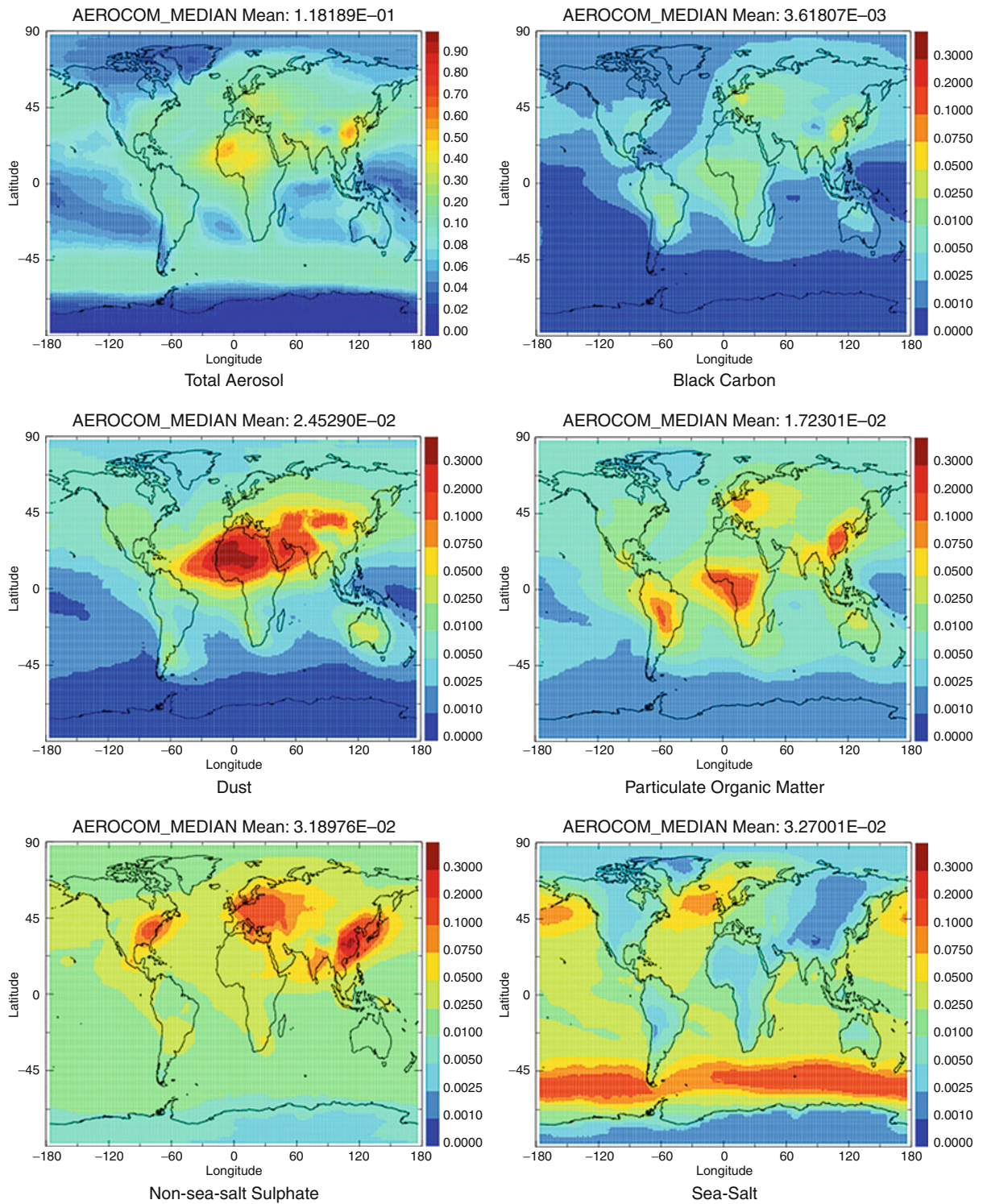
and

$$\tau = \sigma_{\text{ext}} H,$$

where  $\sigma_{\text{ext}}$  is the extinction coefficient of radiation propagating through an aerosol layer of height  $H$ .

In terms of radiative forcing, an aerosol layer can either cause a negative or positive sign change to planetary albedo, leading to a cooling or warming effect.  $\omega$  is the key parameter, along with the albedo of the underlying surface, that determines whether an aerosol layer leads to heating or cooling. Over dark surfaces like the ocean, and regardless of the degree of absorption, albedo will be increased primarily due to the upscatter of incoming radiation. In contrast, over bright surfaces such as snow and desert surfaces, an absorbing component will reduce the amount of surface-reflected radiation radiated back out to space. An additional important factor related to absorbing aerosol layers is that they warm as they absorb radiation, leading to increased atmospheric stability and reduced convection. This effect, it will be seen later, can have important implications for cloud formation.

The global distribution of AOD for total aerosol types and individually for black carbon, particulate organic carbon, sulfate, dust, and sea-salt is shown in Fig. 14. The global average AOD, derived from the medium of the 10 AEROCOM models, is 0.12 (compared to 0.137 derived from satellites and sunphotometers) [57]. Peak AOD values of  $>0.6$  are encountered over Northwest Africa and the Asian continent, and to a lesser degree, over Europe, the USA, and South America. The lowest AOD values of  $\sim 0.01$ – $0.02$  occur over Antarctica. Black carbon is seen to contribute 0.0036 to the global AOD, with hot spots over Europe and China where black carbon contributes 0.05 to AOD. Dust contributes 0.025 to global AOD with



**Aerosol in Global Atmosphere. Figure 14**

Median global distribution of aerosol AOD, total and per aerosol species, derived from 10 AEROCOM models for the year 2000 (Copyright European Geophysical Union 2006. Reprinted from [57])

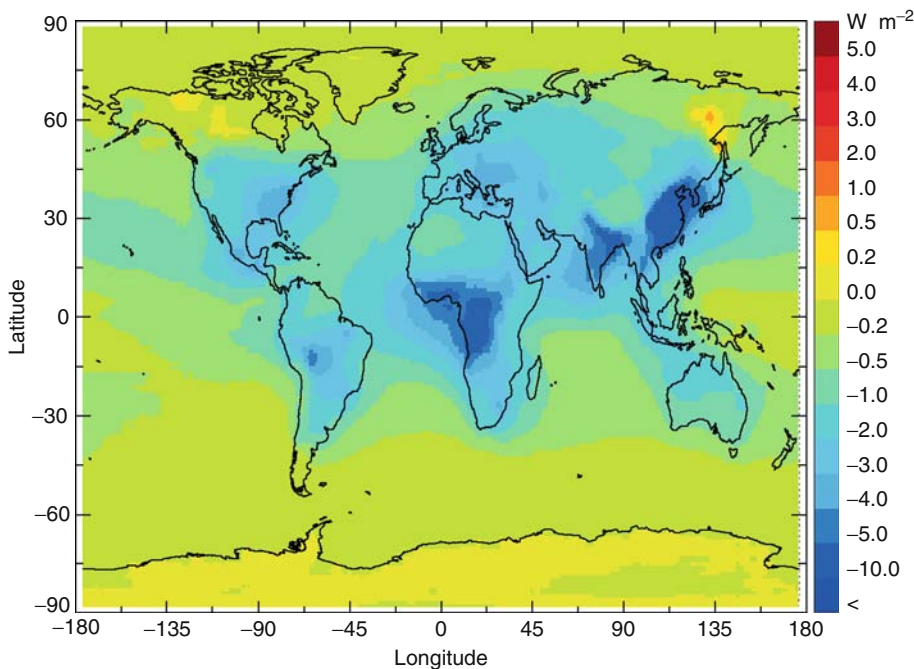
peak contributions of 0.3 arising over Northwest Africa, and 0.2 of Asia. Particulate Organic Matter contributes 0.017 to global AOD, with hot spots in midwest Africa (from biomass burning), South America, some regions in Europe and China contributing AOD greater than 2. Sulfate contributes 0.032 to the global AOD, with Europe, east Asia, and east USA contributing to hot spots with AOD about 0.3. Sea-salt contributes 0.033 to the global AOD, with values up to 0.2 seen in some oceanic regions where wind speeds are generally high (i.e., midlatitudes).

The AEROCOM models predict [76] that, for anthropogenic aerosols, a harmonized estimate for the aerosol direct radiative forcing under all-sky conditions, on a global annual basis, is  $-0.22 \text{ W m}^{-2}$ , ranging from  $+0.04$  to  $-0.41 \text{ W m}^{-2}$ , with a standard deviation of  $\pm 0.16 \text{ W m}^{-2}$ . Anthropogenic nitrate, secondary organic aerosol, and dust were not, however, included in the estimate. By comparison, the IPCC AR4 [100] reported a direct aerosol radiative effect best estimate of  $-0.5 \text{ W m}^{-2}$  with an uncertainty of  $-0.9$  to  $-0.1 \text{ W m}^{-2}$ . The global annual mean aerosol

surface radiative forcing is illustrated in Fig. 15 where it can be seen in the regions with highest aerosol loadings, negative forcings as high as  $-10 \text{ W m}^{-2}$  are encountered. More recent estimates, using a combination of satellite data with a global model [77] suggested a value of  $-0.65 \text{ W m}^{-2}$ . The diversity in model estimates stems more from different model approaches to transport processes, removal mechanisms, and microphysics treatment rather than aerosol emission inventories.

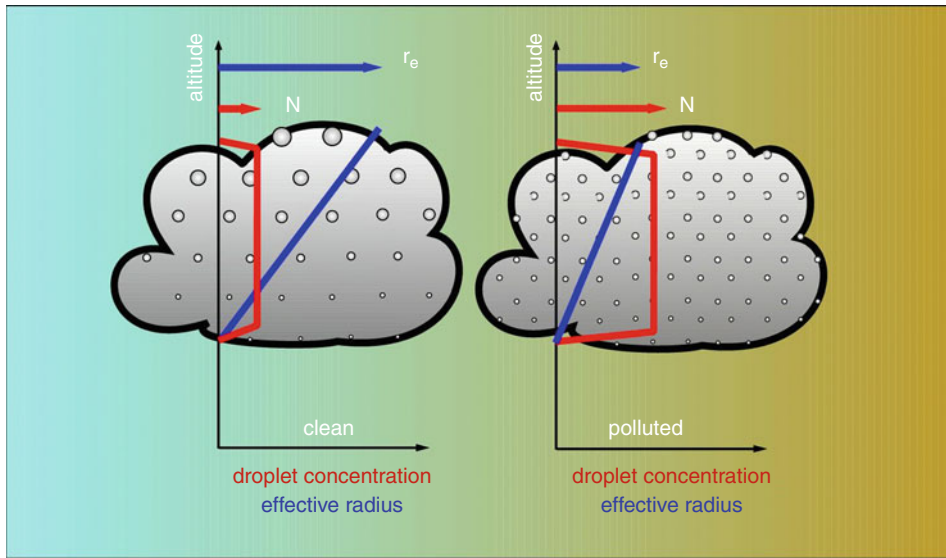
### Indirect Effect

There are a number of indirect aerosol radiative effects whereby aerosol fields modify the cloud microphysics, structure, or lifetime, all of which have albedo impacts. The first indirect effect (typically associated with “warm” or water clouds of fixed liquid water content), results from an increase in cloud condensation nuclei leading to a higher number of smaller cloud droplets which ultimately, through an increase in number of droplets, and the increased scattering associated with



**Aerosol in Global Atmosphere. Figure 15**

Aerosols in the Global Atmosphere Global annual mean aerosol surface radiative forcing from AEROCOM models (Copyright European Geophysical Union 2006. Reprinted from [76])



**Aerosol in Global Atmosphere. Figure 16**

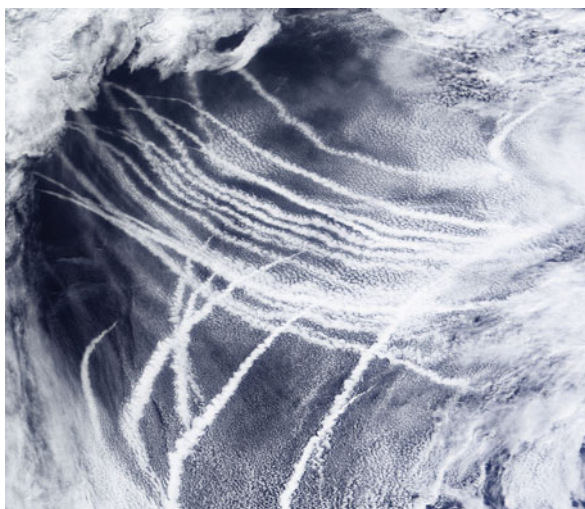
Illustration of changes in cloud microphysics resulting from low (clean) and high (polluted) aerosol fields. For a cloud with the same available liquid water content. With more cloud nuclei available for droplets to form on, the resulting droplets are necessarily smaller in size. The cloud's effective radius is thus smaller and smaller droplets are more efficient scatterers of visible radiation. This, combined with an increase in the number of scattering droplets, results in a more reflective cloud (Courtesy L. Schüller and J.L. Brenguier)

smaller droplets, leads to an albedo increase [72, 74]. Figure 16 illustrates this effect. A cloud is generally formed by rising air parcels which eventually become supersaturated. Cloud droplets are activated at cloud base and continue to grow as the air parcel containing the droplets rise to the top of cloud. Even for a fixed number of droplets, their size and liquid water content increases with increasing height into the cloud. A parameter used to describe an effective radius in such a changing field is the cloud drop effective radius defined as the weighted mean of the size distribution of the cloud droplets. This amounts to the ratio of the volumetric distribution to the surface area distribution – or the ratio of the third moment of the distribution to the second moment. Take the two cases of clouds forming under the same environmental conditions (and hence the same available liquid water content), but under different aerosol fields, one clean containing a low number of cloud nuclei, and the other polluted, containing a high number of cloud nuclei. The same amount of liquid water has to be shared out to different nuclei concentrations entering into cloud

base. The result is that the clean cloud with a low nuclei and droplet concentration ( $N \sim 50 \text{ cm}^{-3}$ ) leads to a relatively large effective radius (of the order of  $15 \mu\text{m}$ ), while the polluted cloud with high nuclei and droplet concentration ( $N \sim 300 \text{ cm}^{-3}$ ) leads to a relatively low effect radius (of the order of  $5 \mu\text{m}$ ). The net effect is a higher cloud albedo seen at cloud top. An excellent example of this effect is seen in Fig. 17, which displays a satellite image of a Pacific Ocean stratocumulus cloud deck off the west US coast. Seen within the cloud deck are numerous tracks more visible than the background cloud. These tracks are called ship tracks and result from ship stack emissions, rich in cloud nuclei, rising into cloud and modifying cloud microphysics as described above. These tracks are easily visible since the tracks have produced regions of increased reflectance relative to the “clean” regions of the cloud. While ship tracks do not necessarily have a climate impact, they are excellent illustrative examples of the first indirect aerosol effect.

The first indirect effect is particularly important over oceans (where surface albedo is low and clouds





**Aerosol in Global Atmosphere. Figure 17**

AVHRR satellite image of a stratocumulus cloud deck off the US Washington coast. The tracks are ship tracks produced by ship stack emissions mixing into cloud, resulting in an increase in cloud droplet concentration and reflectance

are very susceptible to changes in aerosol or nuclei availability) and in regions where persistent and extensive stratocumulus cloud decks occur. It is in these regions that the radiative forcing associated with the indirect effect is at its maximum.

The second indirect effect relates to the cloud lifetime effect [78], and relates to a suppression of precipitation due to the smaller mean droplet sizes. The smaller drops have reduced collision-coalescence efficiency and as a result the onset of precipitation can be delayed or even shut down. Given that precipitation is one important process in shortening the lifetime of a cloud, lack of precipitation can increase cloud lifetime and at times, vertical and horizontal extent. Longer lifetimes or greater extent of clouds effectively increases the albedo since cloud optical depth is always greater than aerosol optical depth.

The semi-indirect effect results from aerosol with a significant absorbing component, which heats the atmospheric layer in which the absorbing aerosol resides [79]. This can inhibit cloud formation by reducing stability and reducing relative humidity through increasing the layer temperature and reducing surface water vapor evaporation. It can also lead to increased

evaporation rates for existing clouds. The sign of the effect is, however, a matter of debate. Some studies [80] support the warming effect of absorbing aerosols through decreased low-level cloud cover and other studies found that cooling effects through increased cloud coverage [81, 82]. The height of the absorbing layer seems to impact on the sign of the effect.

Other effects include the cloud glaciation effect where an increase in anthropogenic ice nuclei can lead to an increase in cloud glaciation [83], the deactivation effect where anthropogenic sulfate coats ice nuclei making them inactive [84], and the thermodynamic effect where smaller droplets freeze less efficiently than larger droplets [85].

From the IPCC AR4, the most recent consensus on the first indirect effect of warm clouds [86] estimated a range between  $-0.22$  and  $-1.85 \text{ W m}^{-2}$ , with a best estimate of  $-0.7 \text{ W m}^{-2}$ . The combined first and second indirect effect was estimated to have a radiative forcing between  $-0.2$  and  $-2.3 \text{ W m}^{-2}$ , illustrating the difficulties in achieving convergence of these models as the level of complexity increases.

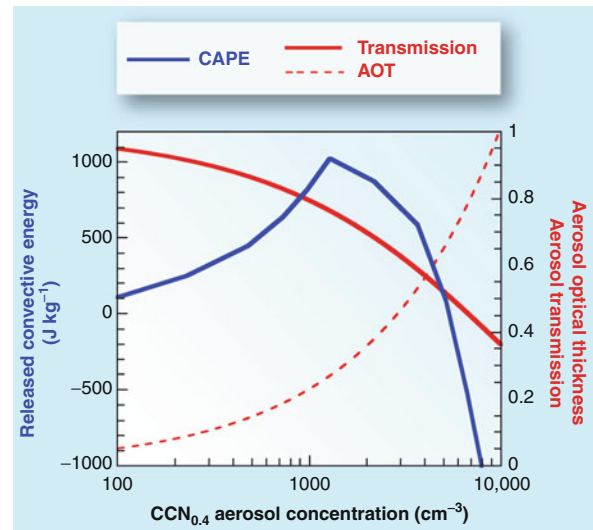
### Aerosol–Cloud–Precipitation Interactions

Although the second indirect effect, through changes in cloud microphysics, has been demonstrated to reduce precipitation with increased anthropogenic aerosol, this effect typically applies to warm stratiform clouds and shallow cumulus clouds. Similarly, the radiative effects generally act to suppress precipitation through decreasing the amount of solar radiation reaching the surface thereby reducing the amount of energy available to evaporate water and energize convective clouds. The radiation not reflected back to space is partly absorbed in the atmosphere in the presence of absorbing carbonaceous aerosol leading to a warming of the air above the surface. This warming stabilizes the lower atmosphere and suppresses convection and the generation of convective clouds.

Some studies have also shown [87] that increases in cloud nuclei can invigorate deep convective rain clouds with warm cloud bases ( $T > 15^\circ\text{C}$ ). In such clouds, the auto-conversion processes (coalescence of droplets into raindrop sizes) are delayed and hence the onset of precipitation in the cloud is also delayed. This results in more water ascending to higher altitudes and colder

temperatures ( $T < 0^\circ\text{C}$ ). Through the delay of precipitation, the additional latent heat release invigorates the cloud. In addition, by not raining early, the condensed water can form ice precipitation particles that release the latent heat of freezing aloft and these reabsorb heat at lower levels where they melt after falling [88]. This process leads to consumption of more convective available potential energy (CAPE) which would then be converted to an equally greater amount of released kinetic energy that could invigorate convection and lead to a greater convective overturning, more precipitation, and deeper depletion of the static instability.

The increased amount of aerosol has two competing effects: one radiative, which suppresses convection (and consequently rainfall) through absorption and warming of atmospheric layers relative to the surface, and the other, which can lead to invigoration of updrafts, leading to an increase in precipitation. These two processes have typically been treated separately leading to opposing views on the impact of aerosols on precipitation. This competing effect was elucidated recently [89] by looking at the combined effect of increasing aerosol and cloud condensation nuclei concentrations on both microphysical and radiative properties. They modeled typical warm tropical convective clouds and found that the microphysical invigoration effect was at a maximum for moderate cloud condensation nuclei concentrations, above which the impact of additional nuclei was reduced. However, as the cloud condensation nuclei concentrations were linearly correlated to AOD, AOD progressively increased with increasing cloud condensation nuclei concentrations which reduced the solar flux which in turn energized convection reaching the surface. As a result, with increasing aerosol loads beyond the optimum, the weakening of the microphysical invigoration is reinforced by the suppressive effect of reduced surface heating. This conceptual model is illustrated in Fig. 18 where it is seen that starting out with clean (low) nuclei concentrations ( $N \sim 100 \text{ cm}^{-3}$ ), the addition of more nuclei has the effect of increasing the released convective energy up to a maximum point corresponding to a cloud condensation nuclei concentration of the order of  $1,200 \text{ cm}^{-3}$ . At this point, AOD is a moderate 0.25. Further increases in cloud condensation nuclei result in a reduction of released convective energy with CAPE dropping rapidly at concentrations



**Aerosol in Global Atmosphere. Figure 18**

Illustration of the relations between the aerosol microphysical and radiative effects. The aerosol optical (dept) thickness (AOT) is assumed to reach 1 at  $\text{CCN}_{0.4} = 10^4 \text{ cm}^{-3}$  (dashed red line), which corresponds to nucleation of 2,000 cloud drops  $\text{cm}^{-3}$ . The related transmission of radiation reaching the surface is shown by the solid red line. The vigor of the convection is shown by the blue line, which provides the released convective available potential energy (CAPE) of a cloud parcel that ascends to the cloud top near the tropopause. Note that a maximum in CAPE occurs at  $\text{CCN}_{0.4} \approx 1,200 \text{ cm}^{-3}$ , which corresponds to the maximum cloud invigoration. The AOT corresponding to the  $\text{CCN}_{0.4}$  at the microphysical optimum is only 0.25. Adding aerosols beyond this point substantially decreases the vigor of the cloud because both microphysical and radiative effects work in the same direction: smaller release of convective energy aloft and less radiative heating at the surface (Reprinted from [89])

over  $5,000 \text{ cm}^{-3}$ , but AOD continues to rise, reaching 1 for extreme nuclei concentrations of  $10,000 \text{ cm}^{-3}$  (corresponding to  $2,000 \text{ cm}^{-3}$  cloud droplets. The slope of the CAPE curve represents the effect of aerosols on precipitation. For increasing CAPE release with increasing aerosol concentrations, an increase in precipitation is expected. This trend reverses for moderate levels of aerosol pollution, above which suppression of precipitation is associated with increasing aerosol concentrations.



## Aerosol–Climate Interactions

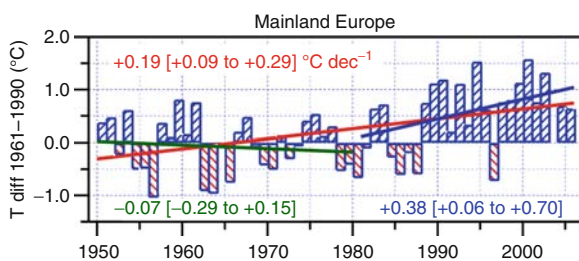
The IPCC AR4 concluded that the combined aerosol radiative forcing was negative and partly offset the warming, or positive forcing, by greenhouse gases. This suggests that aerosols, perhaps, have been masking the true rate of global warming, or the climate-temperature sensitivity to CO<sub>2</sub>-induced global warming. Over the past 40 years, there has been both observed dimming and brightening trends that point to a direct aerosol influence on climate. Global dimming is a term associated with a decadal decrease in surface solar radiation, while global brightening refers to an increase in surface solar radiation. Studies [90–92] have shown a widespread decrease in surface solar radiation at a variety of locations worldwide between 1960 and 1990. Increasing aerosol concentrations associated with increased air pollution over the period are considered responsible for the dimming [93]. Changes in cloud reflectance and cloud amount contribute to the dimming [90]; however, in a particular study over Europe [94], it was concluded that cloud amount could not explain the dimming and that the aerosol direct and indirect effect was the predominant cause for the trend. The period from 1990 to the present shows a reversal of the trend into a brightening trend [95]. The dimming effect appeared to have been masking, or

suppressing, greenhouse warming with reduced or even negative trends on global temperatures over the period as greenhouse gases continued to accumulate. As the trend reversed from dimming to brightening, rapid temperature rises became evident since the mid-1980s. Since the mid-1980s, the decadal rise in temperature has been +0.38°C per decade (Fig. 19), significantly higher than any other period since the preindustrial era [96]. The brightening has been associated with reduced aerosol pollution since the 1980s as developed countries implemented policies to clean up air pollution.

## Aerosols and Air Pollution

Aerosol is one atmospheric constituent regarded as an air pollutant. In Air Quality communities, aerosol particles are termed particulate matter (PM). Aerosol air pollution is typically measured as the total mass less than a particular cutoff size (e.g., 10 µm for PM<sub>10</sub>, 2.5 µm for PM<sub>2.5</sub>, and 1 µm for PM<sub>1</sub>). Air quality is regulated in terms of exposure to a particular PM standard. Air pollution has been a serious problem since the eighteenth century with the invention of the steam engine which increased the amount of coal burning. Prior to that, air pollution was still a problem due to wood and coal burning although not as severe. In 1905, the term “smog” was coined and it described the combination of smoke and fog that was visible in many industrialized cities. London experienced the most frequent and severe smog events resulting in many “excess” deaths. The worst of these occurred in December 1952 when there were 4,000 excess deaths during the London-type event. During this period, smoke mass concentrations reached 4,460 µg m<sup>-3</sup>. This is to be compared with a current EU yearly exposure of 25 µg m<sup>-3</sup> for PM<sub>2.5</sub> and 50 µg m<sup>-3</sup> sustained over 24 h 35 times in 1 year for PM<sub>10</sub>. The events typically occur in winter and during high pressure weather systems.

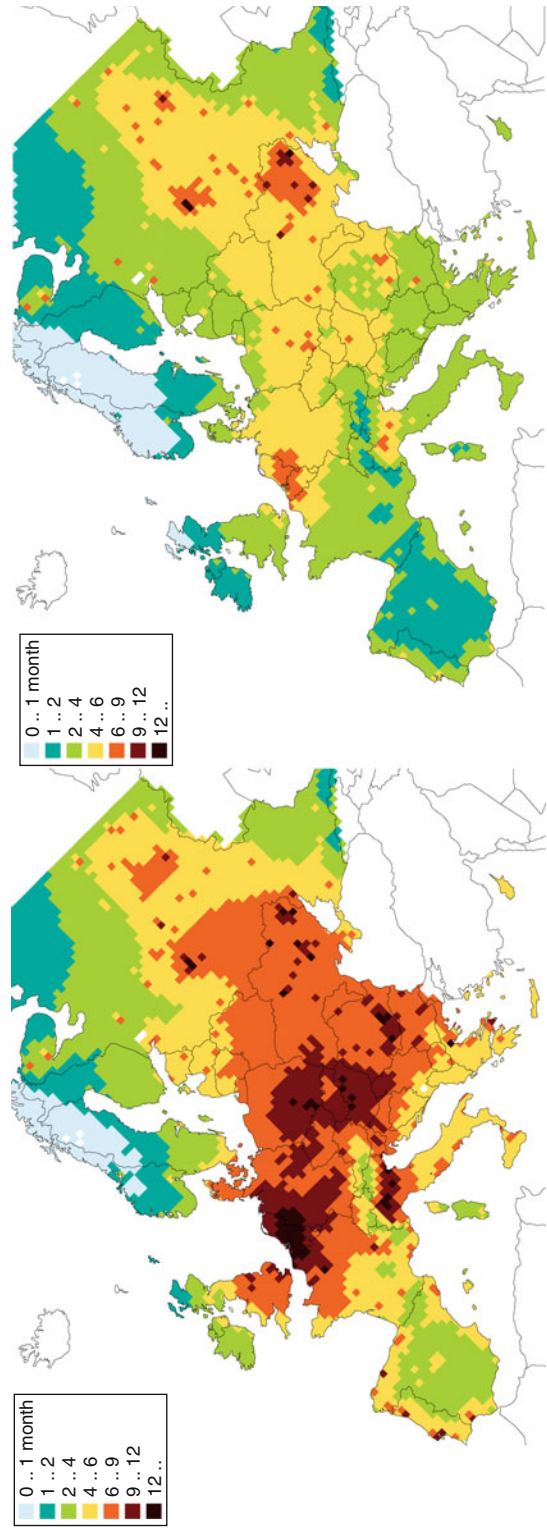
High pressure systems generally are accompanied by colder temperatures resulting in increased coal burning. In addition to the colder temperatures are stable atmospheric boundary layers, which suppress the dispersion of pollution. Further exasperating the situation are very stable surface-mixing layers 100–200 m deep, which trap the pollution and confine the pollution to a thick layer close to the ground.



**Aerosol in Global Atmosphere. Figure 19**

Temperature rise over mainland Europe since 1950. Annual temperature differences for the period 1950 to 2005 with respect to the 1961–1990 climatological mean for mainland Europe (45°–55°N; 5°–15°E). Linear regression lines and decadal trends with 95% confidence interval show the temperature decline from 1950 to 1980 (green), and the temperature rise to be twice as large for the period 1981 to 2005 (blue) than for the whole period 1950 to 2005 (red) (Copyright American Geophysical Union 2009.

Reprinted from [96])

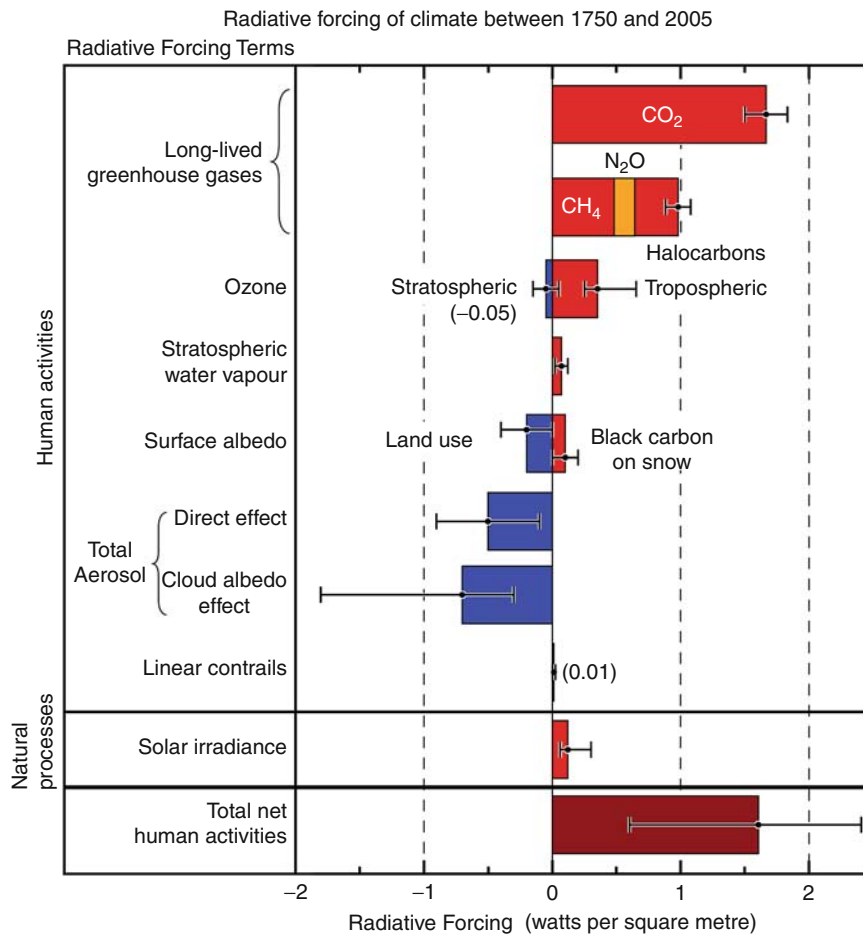


Aerosol in Global Atmosphere. Figure 20

Estimated loss in life expectancy attributable to exposure to fine particulate matter (PM<sub>2.5</sub>) from anthropogenic emissions for the year 2000 (*left*) and projected reduced emissions for the year 2020 (*right*) (Modified from EEA Report No2/2007. Courtesy of Zbigniew Klimont and Markus Amann)

In addition to the London-type smog, there is a chemically produced smog known as photochemical smog. Photochemical smog is produced through gas phase reactions typically involving hydrocarbons, NO, and ozone and is a frequent phenomenon encountered in many megacities such as Los Angeles, Mexico City, Tokyo, Beijing, Johannesburg, and Athens.

The main health impacts of PM relate to respiratory and cardiovascular effects and, additionally, some of the smoke products are considered carcinogenic, all potentially resulting in premature mortality. The European Union's Clean Air for Europe Programme estimated that 348,000 premature deaths occur in Europe due to exposure to PM<sub>2.5</sub>. Figure 20 illustrates the



**Aerosol in Global Atmosphere. Figure 21**

Summary of the principal components of the radiative forcing of climate change. All these radiative forcings result from one or more factors that affect climate and are associated with human activities or natural processes as discussed in the text. The values represent the forcings in 2005 relative to the start of the industrial era (about 1750). Human activities cause significant changes in long-lived gases, ozone, water vapor, surface albedo, aerosols, and contrails. The only increase in natural forcing of any significance between 1750 and 2005 occurred in solar irradiance. Positive forcings lead to warming of climate and negative forcings lead to a cooling. The thin black line attached to each colored bar represents the range of uncertainty for the respective value (Reprinted from the Intergovernmental Panel on Climate Change Assessment Report 4)

estimated losses in life expectancy attributable to exposure to  $\text{PM}_{2.5}$  from anthropogenic emissions in Europe [97]. The data are calculated for emissions for the year 2000 and for targeted emission reductions by the year 2020.

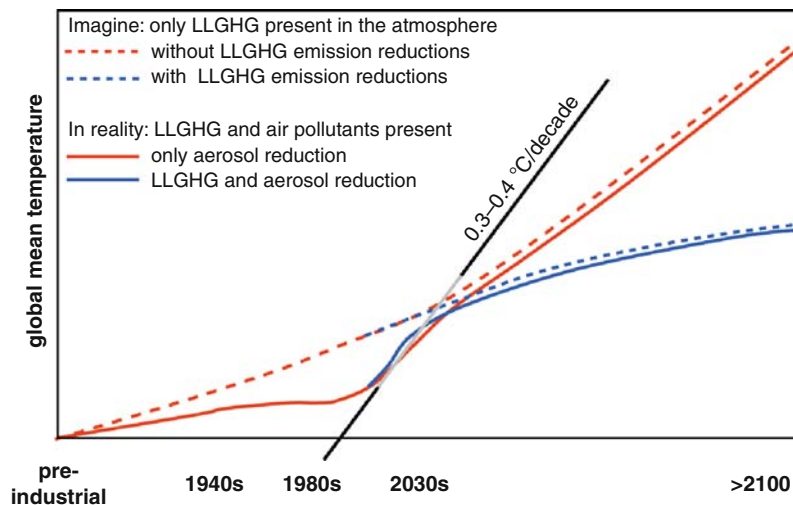
While air pollution, particularly aerosol air pollution, has been steadily reducing in the developed world, it has become an increasing problem in developing countries, not only on urban megacity scales but also on regional and almost hemispheric scales. Intercontinental and hemispheric transport of pollution is now regarded as a serious concern, impacting on local- and regional-scale air quality.

### Future Directions: Interactions Between Aerosol Air Pollution and Climate

Atmospheric aerosol has played an important role in partly offsetting global warming due to greenhouse gases. The IPCC AR4 quantifies the best atmospheric composition component estimates of radiative forcing since 1750 to 2005 (Fig. 21). From this assessment, the

combined direct and indirect effect amounts to  $\sim -1.2 \text{ W m}^{-2}$ , which effectively can be regarded as a reduction in the positive forcing by greenhouse gases to an equivalent amount. The IPCC estimate the net forcing due to anthropogenic activities as  $\sim +1.5 \text{ W m}^{-2}$ . This positive forcing has led to a global temperature increase of the order of  $0.8^\circ\text{C}$ . The European Union, leading the way in climate policy development, have set a long-term (i.e., by year 2050) target of an upper limit of  $2^\circ\text{C}$  to the increase in global temperature.

Until recent times, climate policy has been handled separately to air pollution problems, where in the latter case, adverse health effects from PM and ecosystem damage have been the biggest drivers of policies to improve air quality. What has become evident is that the air pollution policies in recent decades have impacted on global warming in a manner that has accelerated temperature rise as dirty air was cleaned up. The current decadal rate of temperature increase is estimated to be  $0.3\text{--}0.4^\circ\text{C}$  per decade. As a result, by the year 2030, the predicted temperature increase is  $\sim 1.9^\circ\text{C}$  – that is,



**Aerosol in Global Atmosphere. Figure 22**

The figure shows schematically the climate consequences of climate change and air pollution abatement strategies. The only way to reach global temperature stabilization is through LLGHG reductions. Such stabilization would be achieved according to the timescales mainly related to the several decade lifetime of LLGHG in the atmosphere. Reduction of air pollution (including that resulting from LLGHG emission reductions) would have a more immediate impact on global warming. Air pollutants have lifetimes of weeks to months, and the climate system would realize, in about 5 years, 50% of the global warming caused by their reduction [98]. Air pollutant reductions are expected to accelerate the increase in global mean temperature in the short term ( $<30$  years), on the way to eventual long-term ( $>100$  years) climate stabilization (Copyright Elsevier 2009. Reprinted from [99])

almost reaching the long-term target limit. Figure 22 illustrates the problem [99]. If there were only long-lived greenhouse gases (LLGHG), the past temperature increase would have been approximately double the actual increase since the preindustrial era. Without emission reductions, looking into the future, temperature will increase by about 0.2°C per decade. However, taking account of the presence of aerosols, the approximate 50% masking of temperature increase is evident up until the 1990s and then the current day and near future rapid rise in temperature from aerosol emission controls is evident, rapidly approaching the increase seen by treating the system as only containing LLGHG. With both a reduction on aerosol emissions and LLGHG, only then can the dual target of clean air and minimal temperature rise be achieved. Such combined policy development, while in the near term will accelerate temperature rise, in the long term, a sustainable temperature rise can be expected.

The development of sophisticated policy development is underpinned by the development of sophisticated climate and air pollution assessment models. Current climate models do not possess treatments of aerosol emissions, formation, transformations, and radiative effects to a high enough level of sophistication yet. This limitation results from the aerosol life cycle being one of the most complex systems to model and the challenges of modeling the formation of nano-clusters to global-scale radiative impacts and interactions with clouds will remain for some time. Nevertheless, the advances in recent years have been immense, most notably in the current day improved understanding of the highly complex nucleation processes and in the characterization of organic aerosols, their properties, and their life cycle.

## Bibliography

- Seinfeld JH, Pandis SN (1998) *Atmospheric chemistry and physics: from air pollution to climate change*. Wiley, New York/Chichester, air pollution to climate, 1998
- Jaenicke R (1998) Atmospheric aerosol size distribution. In: Harrison RM, van Grieken RE (eds) *Atmospheric particles*. Wiley, Chichester, UK/New York
- Brasseur GP, Prinn RG, Pszenny AAP (eds) (2003) *Tropospheric aerosols. Atmospheric chemistry in a changing world – an integration and synthesis of a decade of tropospheric chemistry research*. Springer-Verlag, Berlin, Heidelberg. ISBN: 3-540-43050-4
- Gillette DA (1979) Environmental factors affecting dust emission by wind erosion. In: Morales C (ed) *Saharan Dust, SCOPE 14*. Wiley, New York, pp 71–91
- Bishop JKB, Davis RE, Sherman JT (25 October 2002) Robotic observations of dust storm enhancement of carbon biomass in the North Pacific. *Science* 298(5594):817–821
- Elbert W, Taylor PE, Andreae MO, Pöschl U (2007) Contribution of fungi to primary biogenic aerosols in the atmosphere: wet and dry discharged spores, carbohydrates, and inorganic ions. *Atmos Chem Phys* 7:4569–4588
- Després VR et al (2007) Characterization of primary biogenic aerosol particles in urban, rural, and high-alpine air by DNA sequence and restriction fragment analysis of ribosomal RNA genes. *Biogeosciences* 4:1127–1141
- Möhler O, DeMott PJ, Vali G, Levin Z (2007) Microbiology and atmospheric processes: the role of biological particles in cloud physics. *Biogeosciences* 4:1059–1071
- Lewis ER, Schwartz SE (2004) *Sea salt aerosol production: mechanisms, methods, measurements and models – a critical review*, vol 152, Geophysical monograph. American Geophysical Union, Washington, DC
- Monahan EC, Spiel DE, Davidson KL (1986) A model of marine aerosol generation via whitecaps and wave disruption. In: Monahan EC, MacNiocaill G (eds) *Oceanic whitecaps and their role in air-sea exchange processes*. Reidel, Dordrecht, pp 167–174
- Blanchard DC (1963) The electrification of the atmosphere by particles from bubbles in the sea. *Progr Oceanogr* 1:71–202
- O'Dowd CD, Lowe JA, Smith MH (1997) Marine aerosol, sea-salt, and the marine sulphur cycle: a short review. *Atmos Environ* 31:73–80
- O'Dowd CD, Facchini MC, Cavalli F, Ceburnis D, Mircea M, Decesari S, Fuzzi S, Yoon YJ, Putaud J-P (2004) Biogenically-driven organic contribution to marine aerosol. *Nature*. doi:10.1038/nature02959
- IEA (1998a) *Energy statistics of OECD countries*. International Energy Agency, Paris
- IEA (1998b) *Energy statistics of non-OECD countries*. International Energy Agency, Paris
- Watson JG et al (1994) Differences in the carbon composition of source profiles for diesel-powered and gasoline-powered vehicles. *Atmos Environ* 28(15):2493–2505
- O'Dowd CD, Aalto P, Hämeri K, Kulmala M, Hoffmann T (2002) Atmospheric particles from organic vapours. *Nature* 416: 497–498
- Oxtoby DW, Kashchiev D (1994) A general relation between the nucleation work and the size of the nucleus in multicomponent nucleation. *J Chem Phys* 100(10):7665–7767
- Riipinen I et al (2007) Connections between atmospheric sulphuric acid and new particle formation during QUEST III–IV campaigns in Heidelberg and Hyytiälä. *Atmos Chem Phys* 7(8):1899–1914
- Weber RJ et al (1997) Measurements of new particle formation and ultrafine particle growth rates at a clean continental site. *J Geophys Res-Atmos* 102(D4):4375–4385



21. McMurry PH, Friedlander SK (1979) New particle formation in the presence of an aerosol. *Atmos Environ* 13(12):1635–1651
22. Kulmala M et al (2004) Formation and growth rates of ultrafine atmospheric particles: a review of observations. *J Aerosol Sci* 35(2):143–176
23. Kulmala M, Kerminen V-M, Anttila T, Laaksonen A, O'Dowd CD (2004) Organic aerosol formation via sulphate cluster activation. *J Geophys Res* 109:D04205. doi:10.1029/2003JD003961
24. Bonn B, Kulmala M, Riipinen I, Sihto SL, Ruuskanen TM (2008) How biogenic terpenes govern the correlation between sulfuric acid concentrations and new particle formation. *J Geophys Res-Atmos* 113:D12209. doi:10.1029/2007JD009327
25. Metzger A, Verheggen B, Dommene J, Duplissy J, Prevota ASH, Weingartner E, Riipinen I, Kulmala M, Spracklen DV, Carslaw KS, Baltensperger U (2010) Evidence for the role of organics in aerosol particle formation under atmospheric conditions. *Proc Natl Acad Sci USA*. doi:10.1073/pnas.0911330107
26. O'Dowd CD, Jimenez JL, Bahreini R, Flagan RC, Seinfeld JH, Pirjola L, Kulmala M, Jennings SG, Hoffmann T (2002) Marine particle formation from biogenic iodine emissions. *Nature* 417:632–636
27. O'Dowd CD, Geever M, Hill MK, Jennings SG, Smith MH (1998) New particle formation: spatial scales and nucleation rates in the coastal environment. *Geophys Res Lett* 25:1661–1664
28. O'Dowd CD (2001) Biogenic coastal aerosol production and its influence on aerosol radiative properties. *J Geophys Res* 106:1545–1550
29. Hegg DA, Hobbs PV (1982) Measurement of sulphate production in natural clouds. *Atmos Environ* 16:2663–2668
30. O'Dowd CD, Lowe JA, Clegg N, Clegg SL, Smith MH (2000) Modelling heterogeneous sulphate production in maritime stratiform clouds. *J Geophys Res* 105:7143–7160
31. Henze DK, Seinfeld JH (2007) Global modeling of secondary organic aerosol formation from aromatic hydrocarbons: high- vs low-yield pathways. *Atmos Chem Phys Discuss* 7:14569–14601
32. Szidat S, Jenk TM, Synal HA, Kalberer M, Wacker L, Hajdas I, Kasper-Giebl A, Baltensperger U (2006) Contributions of fossil fuel, biomass-burning, and biogenic emissions to carbonaceous aerosols in Zurich as traced by  $^{14}\text{C}$ . *J Geophys Res* 111:D07206. doi:10.1029/2005JD006590
33. Kroll JH, Ng NL, Murphy SM, Flagan RC, Seinfeld JH (2005) Secondary organic aerosol formation from isoprene photooxidation under high-NO<sub>x</sub> conditions. *Geophys Res Lett* 32:L18808. doi:10.1029/2005GL023637
34. Decesari S, Mircea M, Cavalli F, Fuzzi S, Tagliavini E, Facchini MC (2007) Source attribution of water-soluble organic aerosol by nuclear magnetic resonance spectroscopy. *Environ Sci Technol* 41:2479–2484. doi:10.1021/es061711l
35. Claeys M et al (2004a) Formation of secondary organic aerosol through photooxidation of isoprene. *Science* 303:1173–1176
36. Claeys M, Wang W, Ion AC, Kourtchev I, Gelencsér A, Maenhaut W (2004b) Formation of secondary organic aerosols from isoprene and gas-phase oxidation products through reaction with hydrogen peroxide. *Atmos Environ* 38:4093–4098
37. Limbeck A, Kulmala M, Puxbaum H (2003) Secondary organic aerosol formation in the atmosphere via heterogeneous reaction of gaseous isoprene on acidic particles. *Geophys Res Lett* 30(19):1996. doi:10.1029/2003GL017738
38. Jang M, Czoschke NM, Lee S, Kamens RM (2002) Heterogeneous atmospheric aerosol production by acid-catalyzed particle phase reactions. *Science* 298:814
39. Kalberer M et al (2004) Identification of polymers as major components of atmospheric organic aerosols. *Science* 303:1659–1662
40. Ervens B, Feingold G, Frost GJ, Kreidenweis SM (2004) A modeling study of aqueous production of dicarboxylic acids: 1. Chemical pathways and speciated organic mass production. *J Geophys Res* 109:D15205. doi:10.1029/2003JD004387, 2004
41. Jimenez JL et al (2009) Evolution of organic aerosols in the atmosphere. *Science* 326:1525. doi:10.1126/science.1180353
42. Andreae MO, Rosenfeld D (2008) Aerosol–cloud–precipitation interactions. Part 1, The nature and sources of cloud-active aerosols. *Earth-Sci Rev*, doi:10.1016/j.earscirev.2008.03.001
43. Andreae MO (1995) Climatic effects of changing atmospheric aerosol levels. In: Henderson-Sellers A (ed) *World survey of climatology*, vol 16, *Future climates of the world*. Elsevier, Amsterdam, pp 341–392
44. Andreae MO, Merlet P (2001) Emission of trace gases and aerosols from biomass burning. *Glob Biogeochem Cycles* 15:955–966
45. Bauer SE et al (2007) Nitrate aerosols today and in 2030: a global simulation including aerosols and tropospheric ozone. *Atmos Chem Phys* 7:5043–5059
46. Bond TC et al (2004) A technology-based global inventory of black and organic carbon emissions from combustion. *J Geophys Res* 109(D14):D14203. doi:10.1029/2003JD003697
47. Guelle W, Schulz M, Balkanski Y, Dentener F (2001) Influence of the source formulation on modeling the atmospheric global distribution of sea salt aerosol. *J Geophys Res* 106(D21):27,509–27,524
48. Gong SL, Barrie LA, Lazare M (2002) Canadian Aerosol Module (CAM): a size segregated simulation of atmospheric aerosol processes for climate and air quality models – 2. Global sea-salt aerosol and its budgets. *J Geophys Res* 107(D24):4779. doi:10.1029/2001JD002004
49. Ito A, Penner JE (2005) Historical emissions of carbonaceous aerosols from biomass and fossil fuel burning for the period 1870–2000. *Glob Biogeochem Cycles* 19:GB2028. doi:10.1029/2004GB002374
50. Junker C, Lioussé C (2006) A global emission inventory of carbonaceous aerosol from historic records of fossil fuel and biofuel consumption for the period 1860–1997. *Atmos Chem Phys Discuss* 4:4897–4927
51. Luo C, Mahowald NM, del Corral J (2003) Sensitivity study of meteorological parameters on mineral aerosol mobilization, transport, and distribution. *J Geophys Res* 108(D15):4447. doi:10.1029/2003JD003483
52. Lioussé C et al (2004) Deriving global quantitative estimates for spatial and temporal distributions of biomass burning

- emissions. In: Granier C, Artaxo P, Reeves CE (eds) Emissions of atmospheric trace compounds. Kluwer, Dordrecht, pp 71–113
53. Penner JE et al (2001) Aerosols, their direct and indirect effects. In: Houghton JT et al (eds) Climate change 2001: the scientific basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 289–348
  54. Stier P et al (2005) The aerosol–climate model ECHAM5-HAM. *Atmos Chem Phys* 5:1125–1156
  55. Zender CS, Miller RL, Tegen I (2004) Quantifying mineral dust mass budgets: terminology, constraints, and current estimates. *Eos Trans AGU* 85(48):509–512
  56. Vignati E, Facchini MC, Rinaldi M, Scannell C, Ceburnis D, Sciare J, Kanakidou M, Myriokefalitakis S, Dentener F, O'Dowd CD (2010) Global scale emission and distribution of sea spray aerosol: sea-salt and organic enrichment. *Atmos Environ* 44:670–677
  57. Kinne S, Schulz M, Textor C, Guibert S, Bauer S, Bernsten T, Berglen T, Boucher O, Chin M, Collins W, Dentener F, Diehl T, Easter R, Feichter J, Fillmore D, Ghan S, Ginoux P, Gong S, Grini A, Hendricks J, Herzog M, Horowitz L, Isaksen I, Iversen T, Koch D, Krol M, Lauer A, Lamarque JF, Lesins G, Liu X, Lohmann U, Montanaro V, Myhre G, Penner J, Pitari G, Reddy S, Seland O, Stier P, Takemura T, Tie X (2006) An AeroCom initial assessment – optical properties in aerosol component modules of global models, /ACP, 6/ 1–22
  58. Dall'Osto M, Ceburnis D, Martucci G, Bialek J, Dupuy R, Jennings SG, Berresheim H, Wenger JC, Healy RM, Facchini MC, Rinaldi M, Giulianelli L, Finessi E, Worsnop D, Ehn M, Mikkilä J, Kulmala M, Sodeau J, O'Dowd CD (2010) Aerosol properties associated with air masses arriving into the North East Atlantic during the 2008 Mace Head EUCAARI intensive observing period: an overview. *Atmos Chem Phys* 10:8413–8435. doi:10.5194/acp-10-8413-2010
  59. Lauer A, Hendricks J (2006) Simulating aerosol microphysics with the ECHAM/MADE GCM – Part II: results from a first multiannual integration. *Atmos Chem Phys* 6:5495–5513
  60. Deshler T (2008) A review of global stratospheric aerosol: measurements, importance, life cycle, and local stratospheric aerosol. *Atmos Res* 90:223–232. doi:10.1016/j.atmosres.2008.03.016
  61. Rosen JM (1971) The boiling point of stratospheric aerosols. *J Appl Meteorol* 10:1044–1046
  62. Carslaw KS, Luo BP, Clegg SL, Peter Th, Brimblecombe P, Crutzen PJ (1994) Stratospheric aerosol growth and HNO<sub>3</sub> and water uptake by liquid particles. *Geophys Res Lett* 21:2479–2482
  63. Hanson DR, Mauersberger K (1988) Laboratory studies of the nitric acid trihydrate: implication for the south polar stratosphere. *Geophys Res Lett* 15:855–858
  64. Deshler T, Larsen N, Weisser C, Schreiner J, Mauersberger K, Cairo F, Adriani A, Di Donfrancesco G, Ovarlez J, Ovarlez H, Blum U, Fricke KH, Dörnbrack A (2003b) Large nitric acid particles at the top of an Arctic stratospheric cloud. *J Geophys Res* 108(D16):4517. doi:10.1029/2003JD003479
  65. Fahey DW et al (2001) The detection of large HNO<sub>3</sub>-containing particles in the winter Arctic stratosphere. *Science* 291: 1026–1031
  66. Larsen N, Knudsen BM, Svendsen SH, Deshler T, Rosen JM, Kivi R, Weisser C, Schreiner J, Mauersberger K, Cairo F, Ovarlez J, Oelhaf H, Schmidt A (2004) Formation of solid particles in synoptic-scale Arctic PSCs in early winter 2002/2003. *Atmos Chem Phys* 4:2001–2013
  67. Köhler H (1936) The nucleus in and the growth of hygroscopic droplets. *Trans Farad Soc* 32:1152–1161
  68. Chylek P, Wong JGD (1998) Erroneous use of the modified Köhler equation in cloud and aerosol physics application. *J Atmos Sci* 55:1473–1477
  69. Pruppacher H, Klett J (1997) Microphysics of clouds and precipitation, atmospheric and oceanographic sciences library; v. 18, Kluwer, Dordrecht; London, includes bibliographical references and index Previous ed.: 1978 "With an introduction to cloud chemistry and cloud electricity"
  70. Rogers RR, Yau MK (1989) A short course in cloud physics, vol. 113, International Series in Natural Philosophy, 3rd edn. Butterworth-Heinemann, Woburn, MA
  71. Twomey S (1977) The influence of pollution on shortwave albedo of clouds. *J Atmos Sci* 34:1149–1152
  72. Twomey S (1977) Atmospheric aerosols. Elsevier, New York
  73. McFiggans G, Artaxo P, Baltensperger U, Coe H, Facchini MC, Feingold G, Fuzzi S, Gysel M, Laaksonen A, Lohman U, Mentel TF, Murphy DM, O'Dowd CD, Snider JR, Weingartner E (2006) The effect of physical and chemical aerosol properties on warm cloud droplet activation. *Atmos Chem Phys* 6: 2593–2649
  74. Vali G (1996) Ice nucleation – a review. In: Kulmala M, Wagner P (eds) Nucleation and atmospheric aerosols. Pergamon, Oxford, pp 271–279
  75. Bohren CF, Huffman D (1983) Absorption and scattering of light by small particles. Wiley, New York
  76. Schulz M, Textor C, Kinne S, Balkanski Y, Bauer S, Bernsten T, Berglen T, Boucher O, Dentener F, Guibert S, Isaksen ISA, Iversen T, Koch D, Kirkevåg A, Liu X, Montanaro V, Myhre G, Penner JE, Pitari G, Reddy S, Seland Ø, Stier P, Takemura T (2006) Radiative forcing by aerosols as derived from the AeroCom present-day and pre-industrial simulations. *Atmos Chem Phys* 6:5225–5246
  77. Bellouin N, Jones A, Haywood J, Christopher SA (2008) Updated estimate of aerosol direct radiative forcing from satellite observations and comparison against the Hadley Centre Climate Model. *J Geophys Res* 113(D10): D10205
  78. Albrecht BA (1989) Aerosols, cloud microphysics, and fractional cloudiness. *Science* 245:1227–1230
  79. Hansen J, Sato M, Ruedy R (1997) Radiative forcing and climate response. *J Geophys Res* 102:6831–6864
  80. Jacobson MZ (2002) Control of fossil fuel particulate black carbon and organic matter, possibly the most effective method of slowing global warming. *J Geophys Resh* 107. doi:10.1029/2001JD001376

81. Menon S, Hansen J, Nazarenko L, Luo Y (2002) Climate effects of black carbon aerosols in China and India. *Science* 297:2250–2252
82. Penner JE, Zhang SY, Chuang CC (2003) Soot and smoke aerosols may not warm climate. *J Geophys Res* 108(21):4657. doi:10.1029/2003JD003409
83. Lohmann U (2002) Possible aerosol effects on ice clouds via contact nucleation. *J Atmos Sci* 59:647–656
84. Girard E, Blanchet J-P, Dubois Y (2004) Effects of arctic sulphuric acid aerosols on wintertime low-level atmospheric ice crystals, humidity and temperature at Alert, Nunavut. *Atmos Res* 73:131–148
85. Rosenfeld D, Woodley WL (2000) Deep convective clouds with sustained supercooled liquid water down to  $-37.5^{\circ}\text{C}$ . *Nature* 405:440–442
86. Forster P, Ramaswamy V, Artaxo P, Berntsen T, Betts R, Fahey DW, Haywood J, Lean J, Lowe DC, Myhre G, Nganga J, Prinn R, Raga G, Schulz M, Van Dorland R (2007) Changes in atmospheric constituents and in radiative forcing. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) *Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK/New York
87. Rosenfeld D, Kaufman YJ, Koren I (2006) Switching cloud cover and dynamical regimes from open to closed Benard cells in response to the suppression of precipitation by aerosols. *Atmos Chem Phys* 6:2503–2513
88. Rosenfeld D (2006) Aerosol–cloud interactions control of earth radiation and latent heat release budgets. *Space Sci Rev* 125:149–157
89. Rosenfeld D, Lohmann U, Raga GB, O'Dowd CD, Kulmala M, Fuzzi S, Reissell A, Andreae MO (2008) Flood or drought: how do aerosols affect precipitation? *Science* 231. doi:10.1126/science.1160606
90. Liepert BG (2002) Observed reductions of surface solar radiation at sites in the United States and worldwide from 1961 to 1990. *Geophys Res Lett* 29. doi:10.1029/2002GL014910
91. Wild M, Ohmura A, Gilgen H, Rosenfeld D (2004) On the consistency of trends in radiation and temperature records and implications for the global hydrological cycle. *Geophys Res Lett* 31:L11201. doi:10.1029/2003GL019188
92. Wild M (2009) Global dimming and brightening: a review. *J Geophys Res* 114:D00D16. doi:10.1029/2008JD011470
93. Stanhill G, Cohen S (2001) Global dimming: a review of the evidence for a widespread and significant reduction in global radiation. *Agric Forest Meteorol* 107:255–278
94. Norris JR, Wild M (2007) Trends in aerosol radiative effects over Europe inferred from observed cloud cover, solar “dimming” and solar “brightening”. *J Geophys Res* 112:D08214. doi:10.1029/2006JD007794
95. Wild M, Gilgen H, Roesch A, Ohmura A, Long CN, Dutton EG, Forgan B, Kallis A, Russak V, Tsvetkov A (2005) From dimming to brightening: decadal changes in surface solar radiation. *Science* 308:847–850
96. Philipona R, Behrens K, Ruckstuhl C (2009) How declining aerosols and rising greenhouse gases forced rapid warming in Europe since the 1980s. *Geophys Res Lett* 36:L02806. doi:10.1029/2008GL036350
97. EEA (2007) *Air pollution in Europe 1990–2004*. EEA Report No 2/2007. European Environment Agency, Copenhagen
98. Ming Y, Ramaswamy V (2009) Nonlinear climate and hydrological responses to aerosol effects. *J Clim* 22:1329–1339
99. Raes F, Seinfeld JH (2009) New directions: climate change and air pollution abatement: a bumpy road. *Atmos Environ*. doi:10.1016/j.atmosenv.2009.06.001
100. IPCC (2007) *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK/New York

---

## Agroecological Basis for Managing Biotic Constraints

CLAUDIO M. GHERSA

IFEVA-CONICET, Depto. de Recursos Naturales y Ambiente, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Agroecosystems and Biological Constraints to Production  
 System Management  
 Conclusions  
 Future Directions  
 Bibliography

### Glossary

**Abiotic factor** A nonliving component of the environment, such as soil, nutrients, light, fire, or moisture.  
**Adaptation** (1) Any aspect of an organism or its parts that is of value in allowing the organism to withstand the conditions of the environment. (2) The evolutionary process by which a species' genome and phenotypic characteristics change over time in response to changes in the environment.

**Agroecology** The science of applying ecological concepts and principles to the design and management of sustainable agroecosystems.

**Agroecosystem** An agricultural system understood as an ecosystem.

**Agroforestry** The practice of including trees in crop- or animal-production agroecosystems.

**Allelopathy** An interference interaction in which a plant releases into the environment a compound that inhibits or stimulates the growth or development of other plants.

**Beneficial insects – arthropods** Beneficial insects are predators, parasites, or competitors of insect pests, helping to regulate pest populations without harm to crops.

**Biomass** The mass of all the organic matter in a given system at a given point in time.

**Biotic factor** An aspect of the environment related to organisms or their interactions.

**Competition** An interaction in which two organisms remove from the environment a limited resource that both require, and both organisms are harmed in the process. Competition can occur between members of the same species and between members of different species.

**Consumer** An organism that ingests other organisms (or their parts or products) to obtain its food energy.

**Decomposer** A fungal or bacterial organism that obtains its nutrients and food energy by breaking down dead organic and fecal matter and absorbing some of its nutrient content.

**Disturbance** An event or short-term process that alters a community or ecosystem by changing the relative population levels of at least some of the component species.

**Diversity** (1) The number or variety of species in a location, community, ecosystem, or agroecosystem. (2) The degree of heterogeneity of the biotic components of an ecosystem or agroecosystem (*see ecological diversity*).

**Domestication** The process of altering, through directed selection, the genetic makeup of a species so as to increase the species' usefulness to humans.

**Dominant species** The species with the greatest impact on both the biotic and abiotic components of its community.

**Ecosystem** A functional system of complementary relations between living organisms and their environment within a certain physical area.

**Generalist** A species that tolerates a broad range of environmental conditions; a generalist has a broad ecological niche.

**Habitat** The particular environment, characterized by a specific set of environmental conditions, in which a given species occurs.

**Herbaceous** Nonwoody.

**Herbivore** An animal that feeds exclusively or mainly on plants. Herbivores convert plant biomass into animal biomass.

**Host** An organism that provides food or shelter for another organism.

**Intercropping** Planting more than one crop in a field using a regular pattern that interleaves each crop in some pattern. A form of polyculture.

**Integrated pest management** Pest control using an array of complementary approaches including natural predators, parasites, pest-resistant varieties, pesticides, and other biological and environmental control practices.

**Legume** A plant in the Leguminosae (Fabaceae) family. Most species in this family can fix nitrogen.

**Microclimate** The environmental conditions in the immediate vicinity of an organism.

**Multi-trophic relationships** The organization of feeding and energy-transfer relationships that determine the path of energy flow through a community or ecosystem that involves organisms of different levels.

**Mycorrhizae** Symbiotic fungal connections with plant roots through which a fungal organism provides water and nutrients to a plant and the plant provides sugars to the fungi.

**Organic matter** Material containing molecules based on Carbon, usually referring to soil organic matter.

**Parasite** An organism that uses another organism for food and thus harms the other organism.

**Parasitism** An interaction in which one organism feeds on another organism, harming (but generally not killing) it.

**Parasitoid** A parasite that feeds on predators or other parasites.

**Patchiness** A measurement of the diversity of successional stages present in a specific area.

**Patchy landscape** A landscape with a diversity of successional stages or habitat types.

**Phenotype** The physical expression of the genotype; an organism's physical characteristics. Phenology is the study of periodic plant and animal life-cycle events and how these are influenced by seasonal and interannual variations in climate

**Polyculture** Cropping systems in which different crop species are grown in mixtures in the same field at the same time.

**Predation** An interaction in which one organism kills and consumes another.

**Predator** An animal that consumes other animals to satisfy its nutritive requirements.

**Primary production** The amount of light energy converted into plant biomass in a system.

**Productivity** The ecological processes and structures in an agroecosystem that enable production.

**Seed bank** The total seed presence in the soil.

**Shifting agriculture** Farming systems that alternate periods of annual cropping with extended fallow periods. "Slash and burn" systems of shifting cultivation use fire to clear fallow areas for cropping.

**Species richness** The number of different species in a community or ecosystem.

**Successional stages** A condition characterized by a particular community of a succession, which is the process by which one community gives way to another.

## Definition of the Subject

Agroecology provides guidelines to develop diversified agroecosystems that take advantage of the effects of the integration of plant and animal biodiversity. From a management perspective, the agroecological objective is to provide balanced environments, sustained yields, biologically mediated soil fertility, and natural pest regulation through the design of diversified agroecosystems and the use of low-input technologies.

## Introduction

Constraints to agricultural production may be classified into four basic categories: abiotic, biotic, socioeconomic, and those related to crop management. The origin and importance of each constraint, their associated losses, and opportunities to alleviate them will

vary for the crop, the input and management levels employed, and the environmental and socioeconomic characteristics of the broader farming system in which the crop is grown. Agronomists and plant protectionists usually address production constraints by focusing on management and input issues to alleviate yield losses caused by particular biotic factors (weed, pest, or disease) that are frequently overestimated, and when added up, exceed crops' yield potential. Moreover, plant breeders, especially with the recent addition of bioengineering and biotechnological tools, have assumed that better varieties alone are able to alleviate the impact of factors curtailing production. From an appreciation of constraints and losses, solutions or opportunities have been proposed, prioritized, and placed into an agenda for action involving technology experimentation, training, socioeconomic and policy support, among other types of intervention. However, there is a need for broader more holistic integration that would jointly assess four broad categories of constraint: biotic, abiotic, management, and socioeconomic. In addition, it is important to focus on how, or through which process constraints to production are generated.

A concern with constraints studies that attempt to average out problems and losses over farms, villages, watersheds, or farming systems is the spatial variation encountered in the types and severity of constraints, in their associated losses, and achieved yields. Variation is often substantial and at a small scale (e.g., parts of the same small field and across field types on a farm). This is especially evident for smallholder farming systems that frequently exhibit historically variable and targeted inputs (e.g., fertilizers) and management superimposed on variability in biophysical factors such as soil types, water availability, weed and pest distribution [1].

Agroecology provides the basic ecological principles needed for studying, designing, and managing agroecosystems that are both productive and that are culturally sensitive, socially just, and economically viable. Instead of focusing on one particular component of the agroecosystem, agroecology emphasizes the interrelatedness of all of its components and the complex dynamics of ecological processes including all environmental and human elements. This approach is based on enhancing the habitat both aboveground and in the soil to produce strong and healthy plants



by promoting beneficial organisms while adversely affecting crop pests (weeds, insects, diseases, and nematodes). From a management perspective, the agroecological objective is to provide balanced environments, sustained yields, biologically mediated soil fertility, and natural pest regulation through the design of diversified agroecosystems and the use of low-input technologies [2, 3].

Agroecologists recognize that intercropping, agroforestry, and other diversification methods mimic natural ecological processes, and that the sustainability of complex agroecosystems lies in the ecological models they follow. By designing farming systems that mimic nature, optimal use can be made of sunlight, soil nutrients, and rainfall [4]. The assumption is that by assembling a functional biodiversity, it is possible to subsidize key processes in the agroecosystem that impact on ecological services, such as the activation of soil biology, the recycling of nutrients, the enhancement of beneficial arthropods and antagonists (Table 1) [5]. Altieri [6] argues that there is evidence supporting that promotion of biodiversity within agricultural systems is the cornerstone strategy of system redesign, since: (1) higher diversity (genetic, taxonomic, structural, and resource) within the cropping system leads to higher diversity in associated biota, (2) increased biodiversity leads to more effective pest control and pollination, and (3) increased biodiversity leads to tighter nutrient cycling.

There is evidence that agroecological diversified agrosystems improve their adaptive capacity and reduces

vulnerability to natural disasters, climate change impacts, and new and emerging environmental and economic system stresses and shocks. This ability of withstanding the impact of factors that may reduce agroecosystem sustainability (systems resilience) can be accomplished through physical, biological, sociocultural, and political means. Aspects such as habitat and crop diversification, in situ conservation of local/indigenous seed and germplasm diversity, maintenance of natural enemies' species diversity, increased carbon sequestration, improved water capture and retention, etc., and diversification of farming systems and local economies; technical, legal, and social support networks for small-scale farmers, rural communities, and indigenous people that reduce socioeconomic vulnerability and strengthen adaptive knowledge processes, etc., can be listed as means to increase sustainability. A study of 208 agroecologically based projects and/or initiatives throughout the developing world documented clear increases in food production over some 29 million hectares, with nearly 9 million households benefiting from increased food diversity and security. Promoted sustainable agriculture practices led to 50–100% increases in per hectare food production (about 1.71 t per year per household) in rain-fed areas typical of small farmers living in marginal environments; that is an area of about 3.58 million hectares, cultivated by about 4.42 million farmers. Such yield enhancements are a true breakthrough for achieving food security among farmers isolated from mainstream agricultural institutions [7].

Crop rotations, polyculture cover crops; intercropping, crop/livestock mixtures are some of the strategies that have been recognized as useful to restore agricultural diversity in both time and space. These strategies exhibit ecological features that have been recognized by different studies. For example, crop rotations incorporate temporal diversity into cropping systems, providing crop nutrients and breaking the life cycles of several insect pests, diseases, and weed life cycles [8]. Polycultures are complex cropping systems in which more than two crop species are planted within sufficient spatial proximity to result in competition or complementation, thus enhancing yields [9, 10]. Intercropping may include trees and animals creating agroforestry systems or mixed crop/livestock mixtures resulting in enhanced complementary relations between components increasing multiple use of the agroecosystem [11]. Moreover, animal integration in

### Agroecological Basis for Managing Biotic Constraints.

**Table 1** Ecological processes to optimize in agroecosystems

|                                                                                                         |
|---------------------------------------------------------------------------------------------------------|
| • Strengthen natural pest-control system                                                                |
| • Decrease toxicity through elimination of agrochemicals                                                |
| • Optimize metabolic function (organic matter decomposition and nutrient cycling)                       |
| • Balance regulatory systems (nutrient cycles, water balance, energy flow, population regulation, etc.) |
| • Enhance conservation and regeneration of soil-water resources and biodiversity                        |
| • Increase and sustain long-term productivity                                                           |

agroecosystems aids in achieving high biomass output and optimal recycling [12]. Cover crops based on the use of pure or mixed stands of legumes or other annual plant species under fruit trees improve soil fertility, enhance biological control of pests, and modify the orchard microclimate [13]. Altieri and Rosset [14] argue that including these, strategies in farming provide for diversified forms of agroecosystems that share the following features:

- (a) Maintain vegetative cover as an effective soil and water conserving measure
- (b) Provide a regular supply of organic matter adding manure, compost, and promotion of soil biotic activity
- (c) Enhance nutrient recycling mechanisms with live-stock systems based on legumes, etc.
- (d) Promote pest regulation enhancing the activity of biological control agents achieved by introducing and/or conserving natural enemies and antagonists

### **Agroecosystems and Biological Constraints to Production**

Plant and animal domestication and technological innovation processes, occurring at different rates depending on ecological and social factors during the expansion of agricultural land use, resulted in the development of sophisticated agricultural systems, which included various types of fallow-crop rotation farming, irrigation, land terracing, soil amendment and fertilizing [15, 16]. In these man-made systems tuned in novel ecosystems, farming became increasingly dependent on labor and capital (technological developments) as intensification maintained or even increased outputs. The population density of domestic animals and humans (slaves and peasants) available to satisfy production needs, such as planting, harvesting, and controlling weeds, became an important factor limiting yields and an important destination for agricultural products. Therefore food and fiber demands were not only driven by the nonagricultural fraction of the society, but were also increasingly demanded as labor populations grew to intensify land production. The social structural changes together with modern technological developments, such as steam and internal combustion engines that occurred since the beginning of industrialization

made possible the replacement of animal and human labor by machinery and traditional by scientific agriculture. Mechanization of farming reduced the time needed to perform activities and was the means for augmenting the loads of energy inputs to the agricultural system in a way that increased productivity as never seen before. Postindustrial intensification that occurred during the twentieth century in most of the more developed countries, known as “The Green Revolution,” became so efficient that the continuous increase in demands for food and fiber by the world’s population could be satisfied despite that during the century’s last decades agricultural land shrank [16].

Food supply and human population growth are related, therefore farming intensification process can be seen as following predator prey pattern of change over time, in which population grows following the increase in food supply, and food supply follows labor/energy inputs to the agroecosystem. However, studies on agricultural development have shown that as society builds up its knowledge base through science and other methods to produce technological innovations, they may further increase yield through intensification of farming practices, even when food demands were well below supply, so intensive farming can be pushed just by technological developments [16]. Paradoxically, once expansion of agriculture stops and technological developments give way to intensification of production, technology innovation may at the same time reestablish the expansion model for agricultural production.

This traditional expansion–intensification model has typified agricultural production systems for the first 75 years of the last century. Gain of agricultural land to the sea by means of polders and drainage systems, the use of river damming to build irrigation oasis into desert lands, transformation of tropical rain forests into coffee, tea, sugar cane, and soybean areas, wetland drainage to produce annual crops are all examples on how technology expanded cropped lands. Perhaps the most interesting example on how technology may help in reestablishing the expansion model is the development of improved cultivars by means of plant breeding programs aimed at increasing crop tolerance to stress factors such as high or low temperatures, drought, pests, and disease and soil salinity. Once these characteristics are bread into crops, there is an

expansion in the environmental ranges in which the crops are planted. Shifting agriculture and expansion of agricultural frontiers enabled the incorporation of highly fertile lands into production. These activities molded self-designed novel ecosystems [17, 18], the agroecosystems, which compensated and many times overcompensated the yield reductions due to soil erosion and increase in problems related to biotic constraints to production (weeds, pests, and diseases) that appeared in sites with longer agricultural histories [19, 20]. When settlements became fixed and expansion rate was reduced, intensification originally depended on the generation of spontaneous technological developments. Later, and especially in modern farming, agricultural and agronomy schools became nodal institutions integrating different and somehow related disciplines to generate the technological solutions to the continually emerging problems for agricultural production. Technological approaches tend to engineer and construct new systems, which use practices, like soil plowing, to destroy the natural vegetation and weeds creating the conditions for crop establishment and growth.

As discussed before, these new systems induced important modifications in the human social and economical structures. Individual and collective reactions to these changes repeatedly appeared in different sites and historical times, sometimes sustained on religious dogmas, others on philosophical views, as naturalism, or on scientific disciplines, such as ecology, that became important for modern societies. Integrated pest management (IPM) and agroecology originated as a reaction challenging the modern industrial farming practices. It emerged as a scientific-based philosophical view that questions the modern expansion–intensification agricultural model and its associated dogma of production at any cost. In the last decades, hundreds of research projects and technological development attempts, aimed at environmentally prone management, have taken place delivering significant information. However, the thrust is still highly technological, and focus is on alleviation or suppression of limiting factors or the symptoms disregarding malfunction of the agroecosystem.

Biotic constraints to agricultural production are broadly identified as weeds (plants), pests (animals), and diseases (fungi and bacteria), and their related biological interactions (namely, competition, herbivory, and predation and parasitism), which cause

reductions in physical yields or yield quality. The prevalent philosophy is that pests, nutrient deficiencies, or other factors are the cause of low productivity, as opposed to the view that pests or nutrients only become limiting if conditions in the agroecosystem are not balanced [21]. This understanding of how production is sustained has diverted agriculturists from realizing that limiting factors only represent symptoms of a more systemic disease inherent to unbalances within the agroecosystem and from an appreciation of the context and complexity of agroecological processes thus underestimating the root causes of agricultural limitations [6].

Pest, disease, and weed problems are strong site- and time-specific imposing a biotic constraint on a crop dimensions. The crop-loss impacts of one organism growing in a location may be entirely different from the losses incurred by the same organism on the same crop in other locales. Further, organism population dynamics, migration, invasion, and damage are driven by local conditions, such as temperature and rainfall, but also by the ecosystem and community complexities. Recognition of the systemic dimensions of biological constraints to production is needed to understand how climate–soil interactions, topography, natural, ecological, and land-use complexity, agricultural history, and cropping practices may determine their occurrence, importance, and frequency [22].

Agroecosystems that have evolved in tropical and subtropical climates have high levels of complexity, including many organisms and multi-trophic interactions. Those that have evolved under temperate or Mediterranean climates may also show complexity, but frequently are structurally simpler and may show strong seasonal variation caused by changes in resource availability during the cold/dry season that stops primary productivity. Topography may cause environmental gradients or patchiness adding complexity at regional and landscape scales. Agricultural land-use history, however, at both regional and landscape scales are important drivers of the agroecosystem complexity and organization of the communities that occur in the different crops. Agriculture in many regions of Europe, Middle-East, Asia, Andean regions in South America, and Central America has been practiced in the same sites through millennia, while in others, covering large

extensions of Australia, South and North America, expansion of agriculture has occurred within the last centuries. During the last decades, land under agriculture in some of these regions with relatively new agroecosystems has shrunk, and intensification occurred in those areas that remained under cropping. In other regions, such as those in the Mato Grosso, Cerrado, and Chaco from Brasil, Paraguay, and Argentina, natural systems have been turned into cropland recently, and this transformation is still happening. These differences in agricultural histories are related to important differences in the organization of the biotic interactions in the agroecosystem. Different selection pressures occurring in agroecosystems select for different adaptations, which include coevolution of crops and organisms that challenge them as well as cropping practices [23]. Therefore, arable land complex communities have evolved in regions with long agricultural histories. These communities have strong interactions among organisms. Communities that were assembled in apparently homogeneous agroecosystems under long periods of monocultures can reach high levels of complexity. For example, Javanese rice fields, which are cropped as simple rice monocultures, can support large numbers of arthropods summed by 765 species that are important for biocontrol [24]. Vegetationally simple cropping systems in the UK have been shown to develop complex belowground biodiversity that may include 100 species of bacteria, 350 species of protozoa, 140 species of nematodes, and 24 distinct types of arbuscular mycorrhizae [25].

Size and variability in the above- and belowground community structures that functionally support the agroecosystem is important to discuss biotic constraints to production and the agroecological basis for managing them. The more complex and diverse ecosystems may sustain greater productivity and stability in ecosystem functions since they are supported by more than one species, and are then less vulnerable to changes in the populations of a particular species due to environmental stress or pest attack. However, complex biotic systems may also deliver more organisms, which may challenge the crop increasing the importance and stability of biotic constraints [26]. There are numerous studies showing that increased plant diversity enhances biological pest control [27, 28], but counter examples also exist where pests or disease levels increase due to

the provision of highly palatable species or changes in canopy microclimate [27–30].

Many of the species originated in regions with long histories of agriculture, and even whole communities have migrated into regions in which crops and agroecosystems have recently expanded. In this way, the flora and fauna of the new agricultural areas of the world such as those in the Argentinean Pampas are dominated by species from the Mediterranean-European agricultural communities. In these new areas, arable land communities may show instability due to the continuous introduction of new organisms and species shifts [18, 31, 32]. Changes in trading systems, land use, and cropping practices, experienced by many regions with long agricultural history, are enhancing instability of biotic communities caused by addition of organisms (*biotic invasions*), extinctions, and species shifts due to change in landscape [33]. The impacts of biotic complexity have significance, especially when cropping systems are simplified to monocultures, as many organisms would impose biotic limitations to production. Diverse cropping systems are typical of many traditional agricultural systems found around the world, particularly in risk-prone environments, but also with long histories of agricultural production [6]. It is also in these regions, primarily in the developing world, where the greatest emphasis on improving diverse production systems through intercropping and agroforestry is found. It is clear that there are trade-offs with more diverse agricultural systems, and that the kind of diversity matters greatly, but the question is how to design diverse systems that can meet multiple goals in an acceptable way [22].

## System Management

As discussed previously, cultural and technological changes that took place during the second half of the twentieth century impacted on land-use patterns and agroecosystem design especially driven by the adoption of industrialized production technologies and reliance on increasing input loads of agrochemicals to alleviate agricultural production constraints. Despite the success in achieving significant increases in global food and fiber yields and yield stability, paradoxically, increasing environmental damage, loss of biodiversity, and associated traditional knowledge were often experienced in the new

agroecosystems. In high-external input agriculture intensification of production due to technological development has increased by augmenting input levels that resulted in increased (biomass) production per unit of land and uniformity of the produce. In many cases, this intensification led to reductions in output stability and resource-use efficiency, and has enhanced overexploitation of natural resource base, and consequently reduced sustainability of agroecosystems [34]. There are many examples of reduction in sustainability by the elimination of growth limitations and of yield reduction factors that induced environmental homogenization and a decrease in genetic variation. Noteworthy are the studies showing that application of biocides to reduce biotic constraints to production impact negatively on organisms that are directly or indirectly beneficial for crop growth [34, 35] and, due to their effects on natural enemies of pests and diseases, further increase the need for biocides control [1].

Agroecological bases for reducing biological constraints to crop production depends on the possibility of managing diversity and disturbance at multiple spatial and temporal scales to use biotic interactions to provide desired agroecosystem service.

### Arable Land Biotic Community Constraints

**Small Scale** Soil disturbance caused by plows and other cultivating devices have been used for centuries in different parts of the world to reduce weed competition as well as for reduction of some pests and diseases. Particular traits are selected by cultural practices in many organisms that survive and reproduce in arable lands and show functional patterns that can be identified [36]. Selection pressures imposed by agricultural practices may be sufficiently strong over extended periods of time so that weed populations can evolve into more competitive populations better adapted to agricultural field conditions than populations from nonagricultural areas [37] or populations from agricultural areas submitted to other type of management. In recent times, management approaches that include reduced tillage have strong impacts on the soil biota of plowed soils. It favors fungal food webs, increases abundance and diversity of predators that can reduce soils weed seed banks, and increases control of pests and diseases. It also changes the environmental conditions reducing germination of seed of many weeds or

changing ontogenic processes and phenology of pests and diseases adapted to plowed-cultivated soil conditions. These effects may alleviate biotic constraints especially caused by weeds; its effects are not stable, and shifts in species happen after this practice is adopted for some time. The design and management of the agroecosystems complexity requires an understanding of general system behavior combined with species- and site-specific knowledge.

The performance of agroecological management could be improved since crops were selected in the ecological conditions likely to occur in the low-input farming. Adaptation to these conditions in which biotic interactions are expected to occur and characteristics such as increased competitiveness to weeds, disease, and pest resistance, enhanced ability to support beneficial rhizosphere microorganisms, and improved capacity to access soil nutrients could all potentially benefit crop growth. Noteworthy is that selecting for such traits and acquiring the necessary knowledge to understand the structure of particular agroecosystem structure dynamics can only be accomplished with long agricultural history to provide for coevolution and the existence of relatively strong interactions. In addition to this, if such a mix of characteristics can be identified, selecting it would be difficult and time consuming. Probably a more practical alternative to adaptation of a single crop to a complex biotic system would be the use of mixtures of cultivars within a field to improve crop production and disease management [38, 39].

Cropping systems may include crop sequences to favor beneficial biotic interactions, such as the use of crops that enhance the populations of beneficial rhizobacteria that reduce diseases and nematodes [40], accompanied by soil cultivation, fallow, and residue management practices aimed at maintaining soil cover and organic matter as well as arbuscular mycorrhiza populations [41] and weed suppression [42]. Competitive crops species or mixtures may be used to reduce the impact of weeds on crop yields. However, holistic management approaches aimed at reducing weeds may integrate the use of other strategies such as manipulating crop seeding density and spatial arrangement, tillage, intercropping, use of allelopathic residues and suppressive mulches, and targeted use of biocontrol agents that reduce weed growth and fecundity during



the growth cycle may be integrated in holistic weed management approaches. Also, various soil management techniques aimed at reducing weed seed survival such as reduced tillage, residue management, and organic matter inputs can be used and integrated with the use of weed-suppressive crop rotations that may also impact on seed survival due to enhanced seed predation or infection by pathogens [39].

In agricultural systems, species are selected to obtain yields but can also be used to manage diversity aiming at encouraging beneficial interactions and minimizing undesired ones [43]. For intercropping systems, species are usually chosen considering species or varieties that differ in rooting patterns, canopy types, phenology, etc., to avoid or reduce negative interference or that have positive interactions such as with the introduction of a legume. Well-designed intercrops can increase overall productivity [10, 44] and potentially reduce risk to farmers. A regional-scale field experiment with rice production in China shows how even a small increase in diversity can have a large impact on system function. In that experiment, simply interplanting two varieties of rice, rather than planting them in separate fields, led to a dramatic reduction in pest problems and pesticide use [45].

**Large Scale** As scale increases, so does the relative importance of species richness because greater numbers of species are needed for the maintenance of ecosystem functions [26]. Loss of diversity in agricultural landscapes has been linked to the disruption of ecological functions such as pest management, pollination services, resistance to plant invasion [46]. Spatial scale is particularly important for both occurrence of pest-related yield constraints to production and for management, because landscape features affect species interactions, microclimates, and weather patterns [24, 46–48]. Landscape structure and dynamics can have notable effects on pests by changing habitat patterns and immigration rates [46, 49–53]. In many regions of the world, agriculture shares space with other land uses forming structural mosaic of habitats with insects and other mobile organisms moving between them [33]. The development of multi-trophic arthropod communities depends on spatial processes (dispersal and foraging) that occur at larger scales than the farm, as well

as temporal processes such as overwintering and reproduction. Habitat fragmentation caused by farming can disrupt both types of process and isolate small natural enemy populations from one another, increasing local extinctions [54].

Several biotic constraints to production could be alleviated if biological control maintained following several strategies that help to enhance indigenous populations of beneficial insects by providing food resources (host prey, pollen and nectar, alternate prey) and shelter for overwintering. Habitat management involves vegetation diversification at multiple scales [55]. Use of insectary plantings or leaving strips of unharvested plants are examples of infield strategies, whereas wildflower borders, grassy buffer strips, windbreaks, and hedgerows are examples of field margin diversification techniques. Larger-scale distribution and connectivity of landscape features such as hedgerows, habitat fragments, and riparian vegetation can also impact levels of biological control as well as provide biodiversity conservation benefits [56]. Interplanting crops with flowering herbaceous plants is promoted as a farmscaping technique since pollen and nectar are essential to the fecundity and longevity of several natural enemy species [57, 58]. Planting of multispecies hedgerows along the edges of farm fields can provide stable habitat and resources for beneficials while fields are bare or crops are young. Biological control may not be enhanced by hedgerows if the availability of pollen and nectar is so high within the hedgerows so that natural enemies do not disperse into adjacent agricultural fields to feed on crop pests [59]; or if the hedgerow attracts new pests, non-pest prey that natural enemies prefer over the crop pest; or top predators that prey on the natural enemies of interest [60–63].

Natural enemy dispersal ranges, which can vary from a few meters to over a kilometer for some parasitoid species [64], will determine the effectiveness of various habitat patterns at enhancing biological control. Blackberry and prune trees provide habitat for alternative hosts of the parasitic wasp, *Anagrus epos*, which later preys upon the vineyard leafhopper pest, *Erythroneura elegantula* [65, 66], but connecting border plantings to infield floral corridors may encourage greater natural enemy movement and biological control in vineyards [63]. Successful conservation

biological control relies upon matching vegetational scale and pattern to the movement range of desired natural enemies in relation to their primary food sources. This requires an expansion beyond habitat management at the field level to incorporate larger landscape patterns and processes, a still relatively unexplored area. In addition to the size and distance between habitat patches, the “matrix” between patches is important for insect movement [67]. Many species that live in habitat patches also utilize resources outside the habitat patch, which is a desirable attribute for biological control allowing for natural enemies to migrate into agricultural fields. Structurally complex landscapes have been found to lead to higher levels of parasitism and lower crop damage [68, 69]; but this is not always the case even within the same region if parasitism rates also depend upon the presence of particular species or plant communities [70, 71].

Various strategies for increasing vegetation diversity within crop fields, including tolerating low levels of weed infestation and intercropping, have been exploited to reduce the density of herbivores attacking crops. Crops within diverse assemblages can be “harder to find and easier to lose” by herbivores, and better protected by natural enemies [72]. Diversified vegetation in field margins and mosaic patches across landscapes can support natural enemies that move into crop fields and provide biocontrol of pest herbivores [71]. Vegetation diversity may also affect management of phytopathogens.

Weed consumption by herbivores could be increased by reducing chemical pesticides to control agricultural pest and weeds, and then adopting resistant or repellent crops to generalist consumers found in agroecosystems. This management practice has been successful with some crop species toxic or deterrent to insects, negatively impacting weeds mainly consumed by herbivores [73]. This also is a good example on how food webs may be reorganized in response to human intervention. Native species of insects, rodents, birds, and other organisms can consume large numbers of weed seeds and promote reductions in requirements for chemical control tactics [74]. Interspersed strips of diverse, phenologically dissimilar crops may better conserve populations of weed seed consumers than crop monocultures [75]. Specialist pathogens have

been used to suppress several weed species in pasture and rangelands, but in annual crops, the difficulty of maintaining appropriate environmental conditions for host infection and the need to achieve rapid weed suppression in a narrow time window has impeded the use of this approach [76].

Meadow strips enhanced pollinator diversity and plant reproductive success in adjacent fields in Switzerland [77]. Similarly, in Costa Rica, proximity to forested areas increased pollination and yield in coffee orchards [78]. Weeds in crop fields supply food for many bird species [79], and hedgerows and other forms of field margin vegetation supply both habitat and food for wildlife species [80].

Timing and intensity of agricultural management such as the way in which soil cropping or irrigating activities are designed may alter habitat distribution and connectivity across the landscape that are very important for maintaining viable populations of different organisms [59, 61, 81]. For example, asynchronous tillage is important for maintaining beetle populations in arable cropland [82] as well as asynchronous flooding is important for sustaining natural enemy populations in rice fields [24]. Initial studies also found that creating a mosaic of crop fields and wetlands in different successional stages had great promise as a strategy for improving waterfowl habitat and sustaining crop production in a multiuse landscape [83]. Interestingly, maintaining rice fields flooded through the winter for waterfowl foraging habitat also provided beneficial agronomic impacts by increasing decomposition of rice straw and reducing grassy weed biomass [84]. Further, numerous studies have shown the interactive effects of landscape complexity and the impacts of agricultural management practices, with more benign practices (such as organic farming) having the greatest effects on increased biodiversity in simple landscapes [46, 85, 86].

## Conclusions

Our understanding of biotic interactions taking place in agroecosystems and how they relate to production constraints is growing rapidly, aided by agroecological approaches and the integration of ecological methodologies and ecologists into agricultural research. Numerous

studies provide data allowing for the characterization of agrosystems that reduce the importance of biological constraints to production. They include great spatiotemporal diversity of crops, discontinuity in monoculture (rotations, early varieties, etc.), a mosaic of small fields to ensure the juxtaposition of cultivated and noncultivated land, the presence of a dominant perennial crop (especially orchards), crops grown with high sowing density to limit weed populations, great genetic diversity in the crops (varieties grown in mixed or alternate rows of crops). Based on them, recommendations can be made concerning the management of cultivated plants and the choice of cropping techniques that consider the spatio-temporal dimension of cropping lands, the composition and abundance of the indigenous flora in and around the fields, soil type, the nature of the environment, and the type of farm.

### Future Directions

It is clear that agroecosystems are self-designed, and their outputs are controlled and depend on the biotic interactions that evolve as they exist. If managing biotic constraints for crop production is to be sustained on agroecological bases, there is a need to understand how agricultural communities are structured taking into consideration the effects of diversity, species composition, and food web structure on ecosystem processes; the impacts of timing, frequency, and intensity of disturbance (at different complexity levels, i.e., whole ecosystem, community, population); and the importance of multi-trophic interactions. All of these aspects have to be observed and explicitly integrated at multiple spatial and temporal scales. The potential for a greater use of agroecological management approaches is high. However, given the variability of biological phenomena, the implementation of the agroecological strategy requires a planned spatiotemporal farm management. However, owing to the nature of these self-assembled ecosystems, there is some inherent unpredictability about responses to different management interventions. Effective synthesis of complex and often apparently contradictory information is still needed. Field-based research that includes monitoring of species performance, along with social learning mediated by farmer-researcher collaborations may help in this task.

### Bibliography

1. Almekinders CJM, Fresco LO, Struik PC (1995) The need to study and manage variation in agro ecosystems. *Neth J Agric Sci* 43:127–142
2. Gliessman SR (1998) *Agroecology: ecological processes in sustainable agriculture*. Ann Arbor Press, Michigan
3. Altieri MA (1999) The ecological role of biodiversity in agroecosystems. *Agric Ecosyst Environ* 74:19–31
4. Pretty JN (1994) *Regenerating agriculture*. Earthscan, London
5. Altieri MA, Nicholls CI (1999) Biodiversity, ecosystem function and insect pest management in agroecosystems. In: Collins WW, Qualset CO (eds) *Biodiversity in agroecosystems*. CRC Press, Boca Raton, pp 69–84
6. Altieri MA (2002) *Agroecology: the science of natural resource management for poor farmers in marginal environments*. *Agric Ecosyst Environ* 93:1–24
7. Pretty J, Hine R (2000) *Feeding the world with sustainable agriculture: a summary of new evidence: final report from "Safe-World" Research Project*. University of Essex, Colchester
8. Sumner DR (1982) Crop rotation and plant productivity. In: Recheigl M (ed) *CRC handbook of agricultural productivity*, vol I. CRC Press, Boca Raton
9. Francis CA (1986) *Multiple cropping systems*. Macmillan, New York
10. Vandermeer J (1989) *The ecology of intercropping*. Cambridge University Press, Cambridge/London
11. Nair PKR (1982) *Soil productivity aspects of agroforestry*. ICRAF, Nairobi
12. Pearson CJ, Ison RL (1987) *Agronomy of grassland systems*. Cambridge University Press, Cambridge
13. Finch CV, Sharp CW (1976) *Cover crops in California orchards and vineyards*. USDA Soil Conservation Service, Washington, DC
14. Altieri MA, Rosset P (1996) Agroecology and the conversion of large-scale conventional systems to sustainable management. *Int J Environ Stud* 50:165–185
15. Chambers R (1983) *Rural development: putting the last first*. Longmans Scientific and Technical/Wiley, Essex, England/New York, p 218
16. Vasey DE (1992) *An ecological history of agriculture 10,000 BC–AD 10,000*. Iowa State University Press, Ames
17. Jorgensen SE, Nielsen SN (1996) Application of ecological engineering principles in agriculture. *Ecol Eng* 7:373–381
18. Hobbs RJ, Salvatore A, Aronson J, Baron JS, Bridgewater P, Cramer VA, Epstein PR, Ewel JJ, Klink CA, Lugo D, Norton DO, Richardson DM, Sanderson ES, Valladares F, Villa M, Zamora R, Zobel M (2006) *Novel ecosystems: theoretical and management aspects of the new ecological world order*. *Glob Ecol Biogeogr* 15:1–7
19. Cox GW, Atkins MD (1979) *Agricultural ecology: an analysis of world food production systems*. W. H. Freeman, San Francisco, p 731
20. Cox CM, Garrett KA, Bockus WW (2005) Meeting the challenge of disease management in perennial grain systems. *Renewable Agric Food Syst* 20:15–24

21. Carrol CR, Vandermeer JH, Rosset PM (1990) *Agroecology*. McGraw-Hill, New York
22. Shennan C (2008) Biotic interactions, ecological knowledge and agriculture. *Philos Trans R Soc Lond B Biol Sci* 363:717–739
23. Ghersa CM, Roush ML, Radosevich SR, Cordray SM (1994) Coevolution of agroecosystems and weed management. *Bioscience* 44:85–94
24. Settle WH, Ariawan H, Astuti EH, Cahyana W, Hakim AL, Hindayana D, Lestari AS (1996) Managing tropical rice pests through conservation of generalist natural enemies and alternative prey. *Ecology* 77:1975–1988
25. Fitter AH, Gilligan CA, Hollingworth K, Kleczkowski A, Twyman RM, Pitchford JW (2005) Biodiversity and ecosystem function in soil. *Funct Ecol* 19:369–377
26. Hooper DU, Chapin FS III, Ewel JJ, Hector A, Inchausti P, Lavorel S, Lawton JH, Lodge DM, Loreau M, Naeem S, Schmid B, Setälä H, Symstad AJ, Vandermeer J, Wardle DA (2005) Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol Monogr* 75:3–35
27. Andow DA (1991) Yield loss to arthropods in vegetationally diverse agroecosystems. *Environ Entomol* 20:1228–1235
28. Andow DA (1991) Vegetational diversity and arthropod population response. *Annu Rev Entomol* 36:561–586
29. Brown BJ, Ewel JJ (1987) Herbivory in complex and simple tropical successional ecosystems. *Ecology* 68:108–116
30. Prieur-Richard AH, Lavorel S, Linhart YB, Dos Santos A (2002) Plant diversity, herbivory and resistance of a plant community to invasion in Mediterranean annual communities. *Oecologia* 130:96–104
31. Ghersa CM, León RJC (1999) Successional changes in the agroecosystems of the rolling pampas. In: Walker LR (ed) *Ecosystems of disturbed ground*. Elsevier, Amsterdam, pp 487–502
32. Radosevich SR, Holt JS, Ghersa CM (2007) *Ecology of weeds and invasive plants*. Wiley, New York
33. Baudry J, Poggio SL, Laurent C (2010) Agricultural landscape changes through globalisation and biodiversity effects. In: Primdahl J, Swaffield S (eds) *Globalisation and agricultural landscapes: change patterns and policy trends in developed countries*. Cambridge University Press, Cambridge, UK, pp 58–70
34. Edwards CA (1990) The importance of integration in sustainable agricultural systems. In: Edwards CA, Lal R, Madden P, Miller RH, House G (eds) *Sustainable agricultural systems*. Soil and Water Conservation Society, Ankeny, pp 249–264
35. Altieri MA (1987) *Agroecology: the scientific basis of alternative agriculture*. Westview, Boulder
36. Freckleton RP, Watkinson AR (2002) Are weed population dynamics chaotic? *J Appl Ecol* 39:699–707
37. Weinig C (2005) Rapid evolutionary responses to selection in heterogeneous environments among agricultural and nonagricultural weeds. *Int J Plant Sci* 166:641–647
38. Mundt CC (2002) Use of multiline cultivars and cultivar mixtures for disease management. *Annu Rev Phytopathol* 40:381–400
39. Shennan C (2008) Biotic interactions in agroecosystems. *Philos Trans R Soc Lond B Biol Sci* 363:717–739
40. Welbaum GE, Sturz AV, Dong ZM, Nowak J (2004) Managing soil microorganisms to improve productivity of agro-ecosystems. *Crit Rev Plant Sci* 23:175–193
41. Gosling P, Hodge A, Goodlass G, Bending GD (2006) *Arbuscular mycorrhizal fungi and organic farming*. *Agric Ecosyst Environ* 113:17–35
42. Anderson RL (2004) Sequencing crops to minimize selection pressure for weeds in the central great plains. *Weed Technol* 18:157–164
43. Moonen AC, Bàrberi P (2008) Functional biodiversity: an agroecosystem approach. *Agric Ecosyst Environ* 127:7–21
44. Mead R, Riley J, Dear K, Singh SP (1986) Stability comparison of intercropping and monocropping systems. *Biometrics* 42: 253–266
45. Zhu Y, Fen H, Wang Y, Li Y, Chen J, Hu L, Mundt CC (2000) Genetic diversity and disease control in rice. *Nature* 406:718–772
46. Tscharnkte T, Klein AM, Krues A, Steffan-Dewenter I, Thies C (2005) Landscape perspectives on agricultural intensification and biodiversity—ecosystem service management. *Ecol Lett* 8:857–874
47. Shennan C, Pisani Gareau T, Sirrine JR (2004) Agroecological approaches to pest management in the US. In: Pretty J (ed) *The pesticide detox, solutions for safe agriculture*. Earthscan, London, pp 193–211
48. Wilby A, Thomas MB (2002) Natural enemy diversity and pest control, patterns of pest emergence with agricultural intensification. *Ecol Lett* 5:353–360
49. Colunga GM, Gage SH, Dyer LE (1998) The insect community. In: Cavigelli MA, Deming SR, Probyn LK, Harwood RR (eds) *Michigan field crop ecology, managing biological processes for productivity and environmental quality*, Michigan State University Extension Bulletin E-2646. Michigan Agricultural Experiment Station, East Lansing, pp 59–70
50. Purtauf T, Roschewitz I, Dauber J, Thies C, Tscharnkte T, Wolters V (2005) Landscape context of organic and conventional farms, influences on carabid beetle diversity. *Agric Ecosyst Environ* 108:165–174
51. Roschewitz I, Hucker M, Tscharnkte T, Thies C (2005) The influence of landscape context and farming practices on parasitism of cereal aphids. *Agric Ecosyst Environ* 108:218–227
52. Roschewitz I, Gabriel D, Tscharnkte T, Thies C (2005) The effects of landscape complexity on arable weed species diversity in organic and conventional farming. *J Appl Ecol* 42:873–882
53. Schmidt MH, Roschewitz I, Thies C, Tscharnkte T (2005) Differential effects of landscape and management on diversity and density of ground-dwelling farmland spiders. *J Appl Ecol* 42:281–287
54. Kalkhoven JTR (1993) Survival of populations and the scale of the fragmented agricultural landscape. In: Bunce RG, Ryszkowski L, Paoletti MG (eds) *Landscape ecology and agroecosystems*. Lewis, Boca Raton, pp 83–90

55. Landis DA, Menalled FD, Costamagna AC, Wilkinson TK (2005) Manipulating plant resources to enhance beneficial arthropods in agricultural landscapes. *Weed Sci* 53: 902–908
56. Burel F, Baudry J, Butet A, Clergeau P, Delettre Y, Le Cœur D, Dubs F, Morvan N, Paillat G, Petit S, Thenail C, Brunel E, Lefeuvre JC (1998) Comparative biodiversity along a gradient of agricultural landscapes. *Acta Oecol* 19:47–60
57. Jervis MS, Kidd MAC, Fitton MD, Huddleson T, Dawah HA (1993) Flower visiting by hymenopteran parasitoids. *J Nat Hist* 27:287–294
58. Idris AB, Grafius E (1995) Wildflowers as nectar sources for *Diadegma insulare* (Hymenoptera, Ichneumonidae), a parasitoid of diamondback moth (Lepidoptera, Yponomeutidae). *Environ Entomol* 24:1726–1735
59. Bugg RL, Ehler LE, Wilson LT (1987) Effect of common knotweed (*Polygonum aviculare*) on abundance and efficiency of insect predators of crop pests. NAL/USDA Report, Beltsville
60. Pollard E (1971) Hedges. VI. Habitat diversity and crop pests, a study of *Brevicoryne brassicae* and its syrphid predators. *J Appl Ecol* 8:751–780
61. Bugg RL, Pickett CH (1998) Introduction, enhancing biological control—habitat management to promote natural enemies of agricultural pests. In: Pickett CH, Bugg RL (eds) *Enhancing biological control: habitat management to promote natural enemies of agricultural pests*. The Regents of the University of California, Berkeley, pp 1–23
62. Rosenheim JA, Limburg DD, Colfer RG (1999) Impact of generalist predators on a biological control agent, *Chrysoperla carnea*, direct observations. *Ecol Appl* 9:409–417
63. Nicholls CI, Parrella M, Altieri MA (2001) The effects of a vegetational corridor on the abundance and dispersal of insect biodiversity within a northern California organic vineyard. *Landscape Ecol* 16:133–146
64. Corbett A (1998) The importance of movement in the response of natural enemies to habitat manipulation. In: Pickett CH, Bugg RL (eds) *Enhancing biological control, habitat management to promote natural enemies of agricultural pests*. University of California Press, Berkeley, pp 25–48
65. Douthett RL, Nakata J (1973) The *Rubus* leafhopper and its egg parasitoid, an endemic biotic system useful in grape pest management. *Environ Entomol* 2:381–386
66. Murphy BC, Rosenheim JA, Granett J, Pickett CH, Dowell RV (1998) Measuring the impact of a natural enemy refuge, the prune tree/vineyard example. In: Pickett CH, Bugg RL (eds) *Enhancing biological control, habitat management to promote natural enemies of agricultural pests*. University of California Press, Berkeley, pp 297–309
67. Ricketts TH, Daily GC, Ehrlich PR, Michener CD (2004) Economic value of tropical forest to coffee production. *Proc Natl Acad Sci USA* 101:12579–12582
68. Thies C, Tscharntke T (1999) Landscape structure and biological control in agroecosystems. *Science* 285:893–895
69. Pullaro TC, Marino PC, Jackson DM, Harrison HF, Keinath AP (2006) Effects of killed cover crop mulch on weeds, weed seeds, and herbivores. *Agric Ecosyst Environ* 115:97–104
70. Menalled FD, Marino PC, Gage SH, Landis D (1999) Does agricultural landscape structure affect parasitism and parasitoid diversity? *Ecol Appl* 9:634–641
71. Landis DA, Wratten SD, Gurr GA (2000) Habitat management to conserve natural enemies of arthropod pests in agriculture. *Annu Rev Entomol* 45:175–201
72. Gurr GM, Wratten SD, Altieri MA (eds) (2004) *Ecological engineering for pest management: habitat manipulation for Arthropods*. CSIRO, Collingwood, p 244
73. Liebman M, Davis AS (2000) Integration of soil, crop and weed management in low-external-input farming systems. *Weed Res* 40:27–47
74. Westerman P, Liebman M, Menalled FD, Heggenstaller AH, Hartzler RG, Dixon PM (2005) Are many little hammers effective?—velvetleaf (*Abutilon theophrasti*) population dynamics in two- and four-year crop rotation systems. *Weed Sci* 53:382–392
75. Heggenstaller AH, Menalled FD, Liebman M, Westerman PR (2006) Seasonal patterns in post-dispersal seed predation of *Abutilon theophrasti* and *Setaria faberi* in three-cropping systems. *J Appl Ecol* 43:999–1010
76. Ghorbani R, Leifart C, Seel W (2005) Biological control of weeds with antagonistic plant pathogens. *Adv Agron* 86: 191–225
77. Albrecht M, Duelli P, Muller C, Kleijn D, Schmid B (2007) The Swiss agri-environment scheme enhances pollinator diversity and plant reproductive success in nearby intensively managed farmland. *J Appl Ecol* 44:813–822
78. Ricketts TH, Daily GC, Ehrlich PR, Fay JP (2001) Countryside biogeography of moths in a fragmented landscape, biodiversity in native and agricultural habitats. *Conserv Biol* 15: 378–388
79. Marshall EJP, Brown VK, Boatman ND, Lutman PJW, Squire GR, Ward LK (2003) The role of weeds in supporting biological diversity within crop fields. *Weed Res* 43:77–89
80. Marshall EJP, Moonen AC (2002) Field margins in northern Europe: their functions and interactions with agriculture. *Agric Ecosyst Environ* 89:5–21
81. Baudry J, Papy F (2001) The role of landscape heterogeneity in the sustainability of cropping systems. In: Nösberger J, Geiger HH, Striuk PC (eds) *Crop science – progress and prospects*. CABI, Oxon
82. Holland JM, Thomas CFG, Birkett T, Southway S, Oaten H (2005) Farm-scale spatiotemporal dynamics of predatory beetles in arable crops. *J Appl Ecol* 42:1140–1152
83. Shennan C, Bode CA (2002) Integrating wetland habitat with agriculture. In: Jackson LL, Jackson D (eds) *The farm as a natural habitat*. Island, Washington, DC, pp 189–204
84. van Groenigen JW, Burns EG, Eadie JM, Horwath WR, van Kessel C (2003) Effects of foraging waterfowl in winter flooded rice fields on weed stress and residue decomposition. *Agric Ecosyst Environ* 95:289–296



85. Benton TG, Vickery JA, Wilson JD (2003) Farmland biodiversity: is habitat heterogeneity the key? *Trends Ecol Evol* 18:182–188
86. Bengtsson J, Ahnstrom J, Weibull AC (2005) The effects of organic agriculture on biodiversity and abundance, a meta-analysis. *J Appl Ecol* 42:261–269

## Agronomic Interactions with CO<sub>2</sub> Sequestration

RATTAN LAL

OARDC/FAES Kottman Hall, Carbon Management and Sequestration Center, The Ohio State University, Columbus, OH, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Soil Carbon Budget

Hidden Carbon Costs

Farming Carbon

Ecosystem Services and Soil Carbon Sequestration

Carbon Sequestration and Agronomic Production

Future Directions

Conclusions

Bibliography

### Glossary

**Carbon sequestration** Transfer of atmospheric CO<sub>2</sub> into long-lived reservoirs

**Mean residence time** The duration during which CO<sub>2</sub>-C resides in a specific pool (pool)

**Soil quality** Capacity of a soil to perform ecosystem services

### Definition of the Subject

Atmospheric concentration of carbon dioxide (CO<sub>2</sub>) can be stabilized, if not reduced, by reducing anthropogenic emissions, sequestering emissions, or both. Emission reduction implies identifying and using no-carbon (C) or low-C energy sources such as alternatives to fossil fuel including wind, solar, hydro,

geo-thermal, biofuels, etc. Reductions in gaseous emissions can also be achieved by enhancing the energy use efficiency. In agronomic systems, involving practices to raise crops and livestock, enhancing energy efficiency implies a range of practices which increase agricultural productivity per unit input of energy-based resources (i.e., fertilizers, pesticides, irrigation). Another strategy is elimination or reduction in frequency and intensity of tillage operations and converting plow-based tillage to no-till farming or conservation agriculture. Use of solar and wind energy for grain drying, water pumping, and heating buildings can also reduce emissions.

Agricultural soils and ecosystems can also be used for sequestration of atmospheric CO<sub>2</sub> by enhancing photosynthesis, increasing net primary productivity (NPP), and converting some of the NPP into stable biomass (forest products) and the soil C pool. The biomass in forest, with a long mean residence time (MRT), has two distinct but related components: the above-ground biomass and the below-ground biomass. The above-ground biomass can be alive or the detritus material. The photosynthates transferred deep into the subsoil through a tap root system have a long MRT. Agroforestry systems, growing crops and raising livestock in combination with perennials (tree species), can enhance the ecosystem C pool by increasing both biomass-C and soil-C components.

Sequestering C in soil entails increasing soil organic C (SOC) pool and also the soil inorganic C (SIC) pool. The SOC pool has three related components: labile/active pool, intermediate pool, and the passive/recalcitrant pool with MRT of < year, <decades to a century, and several millennia, respectively. The goal is to transfer the labile pool into intermediate and preferably passive pools through conversion to conversion tillage, use of manure/compost and other biosolids (biochar), and complex cropping/farming systems. Sequestration of SIC occurs through formation of secondary carbonates. In irrigated systems, however, leaching of bicarbonates is also an important mechanism of SIC transfer into the groundwater. Agronomic practices strongly interact with strategies of reducing emissions and sequestering CO<sub>2</sub> in soils and biota. Therefore, the strategy of agronomic management is to identify the interactive agronomic practices which enhance ecosystem (biotic and pedologic) C pools.

## Introduction

Limiting global warming to 2°C necessitates identification and adoption of diverse strategies which reduce the net anthropogenic emissions of greenhouse gases (GHGs). Agriculture, hitherto a major source of GHGs, is an important industry which can be made more energy-efficient. Importance of improving agricultural production, by as much as 70% between 2010 and 2050, cannot be overemphasized in view of the need to reduce hunger, malnutrition, and poverty. Although a major challenge, making agriculture emission-neutral is a prudent long-term strategy. It is in this context that understanding of the agronomic interactions with CO<sub>2</sub> sequestration in soil is important.

Agronomic systems are defined as site-specific management of soils and crops on the basis of eco-regional and physiographic characteristics, and in the context of socioeconomic and policy environments. These systems are strong determinants of agricultural production, sustainable use of resources, and their environmental impact. Pertinent examples of the environmental impacts of agronomic systems include accelerated soil erosion, nonpoint source pollution, emission of GHGs, and alterations in ecosystem and soil C pools. Most soils of agricultural ecosystems (i.e., croplands, grazing lands) are depleted of their SOC pool. In comparison with the natural/climax vegetation, the remaining SOC pool in agricultural soils may be 20–50% of the antecedent pool under undisturbed conditions [1]. The magnitude of loss is large in soils characterized by high than low antecedent pool, coarse than fine texture, in warm than cold climates, and in degraded/eroded than favorable quality. Furthermore, the magnitude of loss is more from soils managed by extractive farming than of intensive/science-based agriculture. The historic loss is also equivalent to the potential soil C sink capacity, a part of which can be realized through conversion to a restorative land use and adoption of recommended management practices (RMPs).

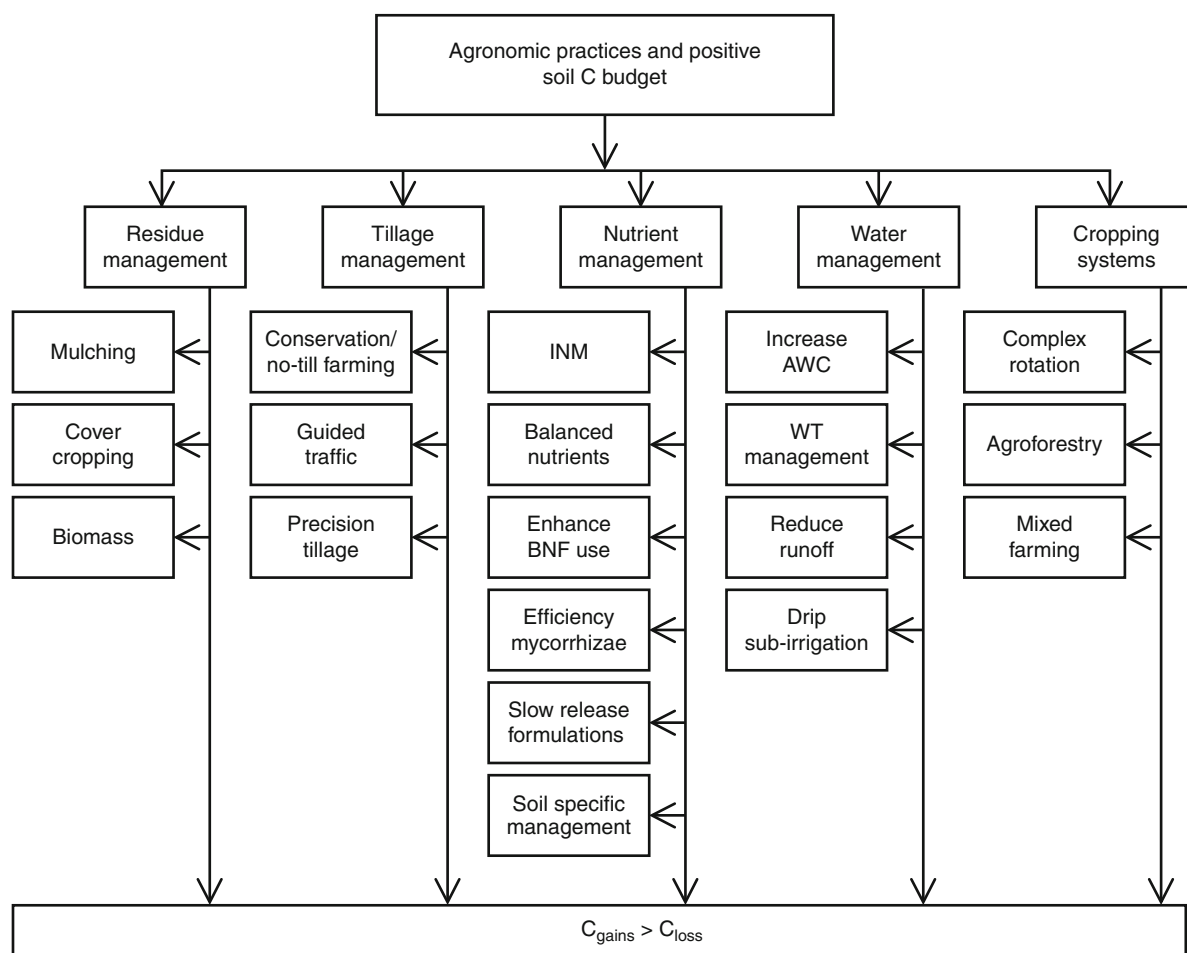
Sequestration of CO<sub>2</sub> in agroecosystems implies transfer of atmospheric CO<sub>2</sub> into biota and soils through photosynthesis and NPP in a manner such that the biomass-C is neither readily nor immediately returned back to the atmosphere. The strategy is to

enhance the MRT of C in biota and soils. Principal determinants of MRT comprise of a wide range of factors including soil processes and plant/biomass characteristics. Important among soil processes are formation of: (1) stable micro-aggregates, (2) organo-mineral complexes including absorption on clay surfaces, and (3) recalcitrant organic polymers involving physical, chemical, and biological protection of soil organic matter (SOM). Another process of physical protection is transfer of SOM into the subsoil by illuviation as dissolved organic (DOC) such that it is away from the surface layer prone to accelerated erosion, intense mineralization, and other natural and anthropogenic perturbations. Important among plant characteristics which increase MRT are: (1) a deep root system and (2) a high concentration of recalcitrant compounds. This chapter is aimed at describing the processes and practices which moderate the agronomic interactions with CO<sub>2</sub> sequestration in soil as a recalcitrant humus of a long MRT.

## Soil Carbon Budget

The strategy is to create a positive SOC budget so that input of biomass-C exceeds the losses. Important among agronomic practices which create a positive C budget (Fig. 1) are mulch farming, no-till/conservation agriculture, integrated nutrient management (INM) including slow release formulation of chemical fertilizers and biofertilizers, conservation and management of soil water to reduce losses by surface runoff and evaporation and increase soil-water storage, and use of complex cropping/farming systems including agroforestry and mixed farming systems. The strategy is to replace extractive farming practices, which deplete soil fertility and SOC pool, by science-based agriculture involving the widespread adoption of RMPs. Some RMPs outlined in Fig. 1 are generic, and no one practice is universally applicable because of the extreme diversity of soil types, ecoregions, and the human dimensions related to socioeconomic and political consideration. Site/soil-specific validation and adaption through fine-tuning of RMPs is essential.

While increasing the input of biomass-C, it is equally important to reduce its losses. Agronomic/soil

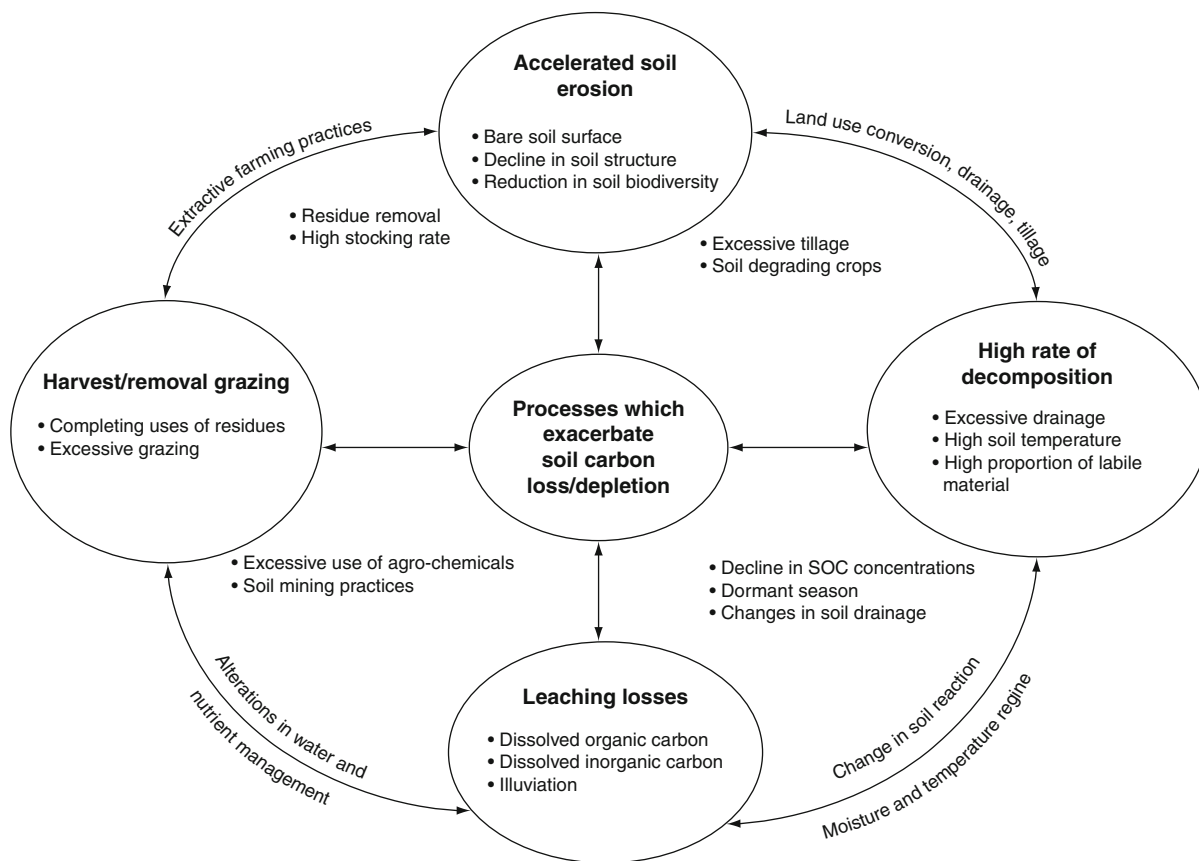


**Agronomic Interactions with CO<sub>2</sub> Sequestration. Figure 1**

Agronomic interactions which cause positive carbon budget (*INM* integrated nutrient management, *AWC* available water capacity, *WT* water table, *BNF* biological nitrogen fixation)

processes which deplete the SOC pool are outlined in Fig. 2. The SOC pool is strongly depleted by accelerated soil erosion. The preferential/selective removal of SOC by runoff and erosion, as indicated by an enrichment ratio of C in sediments ranging from 5 to 30 depending on soil/land and climate, is attributed to: (1) low density of SOC, (2) high concentration (stratification) in the surface layer, and (3) absorption on clay and thus removal along with the clay particles. Soil erosion is exacerbated by plowing, residue removal, uncontrolled and excessive grazing, and management practices which degrade soil structure and accentuate its vulnerability to climatic erosivity.

The SOC pool is also depleted by increase in the rate of mineralization. The latter increases with increase in soil temperature and changes in soil moisture regime. Conversion of natural to managed/agricultural ecosystems alters both the soil temperature and moisture regimes and accentuates the rate of mineralization. The latter, being a biochemical reaction, is approximately doubled with every 10°C increase in temperature (Vant Hoff rule). There is an optimal soil moisture regime for the mineralization/decomposition. Water table management and drainage of excessively wet soils increase the rate of mineralization. Excessive wetness increases methanogenesis and denitrification with an



**Agronomic Interactions with CO<sub>2</sub> Sequestration. Figure 2**  
Agronomic soil processes which deplete the soil carbon pool

attendant increase in emissions of CH<sub>4</sub> and N<sub>2</sub>O. In contrast, supplemental irrigation in arid and semi-arid climates can also accentuate mineralization and denitrification. In general, therefore, emissions of GHGs may be more from agricultural soils than those under natural ecosystems.

### Hidden Carbon Costs

Most agronomic inputs, especially in intensively managed systems, are based on use of fossil fuel combustion. Important among these are tillage operations, harvesting, drying, application of fertilizers, and other chemicals (pesticides) and irrigation. The hidden C cost (HCC) of these inputs are listed in Table 1. Agronomic practices with a significant input of HCC are tillage systems, and agricultural chemicals. Thus,

a complete Life Cycle Analysis (LCA) is needed to determine the ecosystem C budget and assess the net C gains. The latter can be described by Eq. 1:

$$\text{Net C gain} = C_{\text{input}} - (C_{\text{loss}} + \text{HCC}) \quad (1)$$

$C_{\text{input}}$  includes biomass addition such as shoot, leaves, detritus materials, roots, compost, manure, deposition through wind and water, etc.  $C_{\text{loss}}$  occurs mainly through erosion, decomposition, and leaching. Principal components of HCC are the energy-based inputs. Quantification of each of these components is essential to conducting LCA for specific soil, crop, ecoregion, and other site-specific factors.

It is often argued that HCC should not be deducted from the gross C gains because agronomic inputs (i.e., fertilizers, tillage, pesticides, manure, irrigation, and harvesting) are not used for C sequestration but for

**Agronomic Interactions with CO<sub>2</sub> Sequestration.****Table 1** Hidden carbon costs of agronomic practices  
(Adapted from [2])

| Agronomic practice |                                   | Carbon cost |
|--------------------|-----------------------------------|-------------|
| I.                 | <i>Tillage (kg CE/ha)</i>         |             |
|                    | Moldboard plowing                 | 15.2 ± 4.1  |
|                    | Chisel plowing                    | 7.9 ± 2.3   |
|                    | Disking                           | 8.3 ± 2.5   |
|                    | Cultivation                       | 4.0 ± 1.9   |
|                    | Rotary hoeing                     | 2.0 ± 0.9   |
| II.                | <i>Fertilizers (kg CE/kg)</i>     |             |
|                    | Nitrogen                          | 1.3 ± 0.3   |
|                    | Phosphorus                        | 0.2 ± 0.06  |
|                    | Potassium                         | 0.15 ± 0.06 |
|                    | Lime                              | 0.16 ± 0.11 |
| III.               | <i>Pesticides (kg CE/kg)</i>      |             |
|                    | Herbicides                        | 6.3 ± 2.7   |
|                    | Insecticides                      | 5.1 ± 3.0   |
|                    | Fungicides                        | 3.9 ± 2.2   |
| IV.                | <i>Irrigation (kg CE/ha/year)</i> |             |
|                    | Surface                           | 9.4–24.6    |
|                    | Sprinkler                         | 16.3–121.3  |
|                    | Trickle                           | 84.9        |

achieving food security to meet the food and other demands (i.e., feed, fiber, and raw materials) of the growing population. If HCCs are also considered, land managers/farmers are not adequately rewarded by payments through C trading.

**Farming Carbon**

The term “farming carbon” implies growing/increasing C pool in soils and biota (trees) of managed ecosystems (agriculture, forestry, urban lands, wetlands) so that any increase in the ecosystem C pool can be traded in the market as a farm produce. Agronomic interactions, practices which enhance soil and the biotic C pools, are strong determinants of the C gains (or losses) from the ecosystem. Changes in the ecosystem C pool must be

monitored by a standardized procedure so that the data are credible, reproducible, and verifiable by other procedures. A protocol for measurement, monitoring, and verification (MMV) is essential to implement C trading.

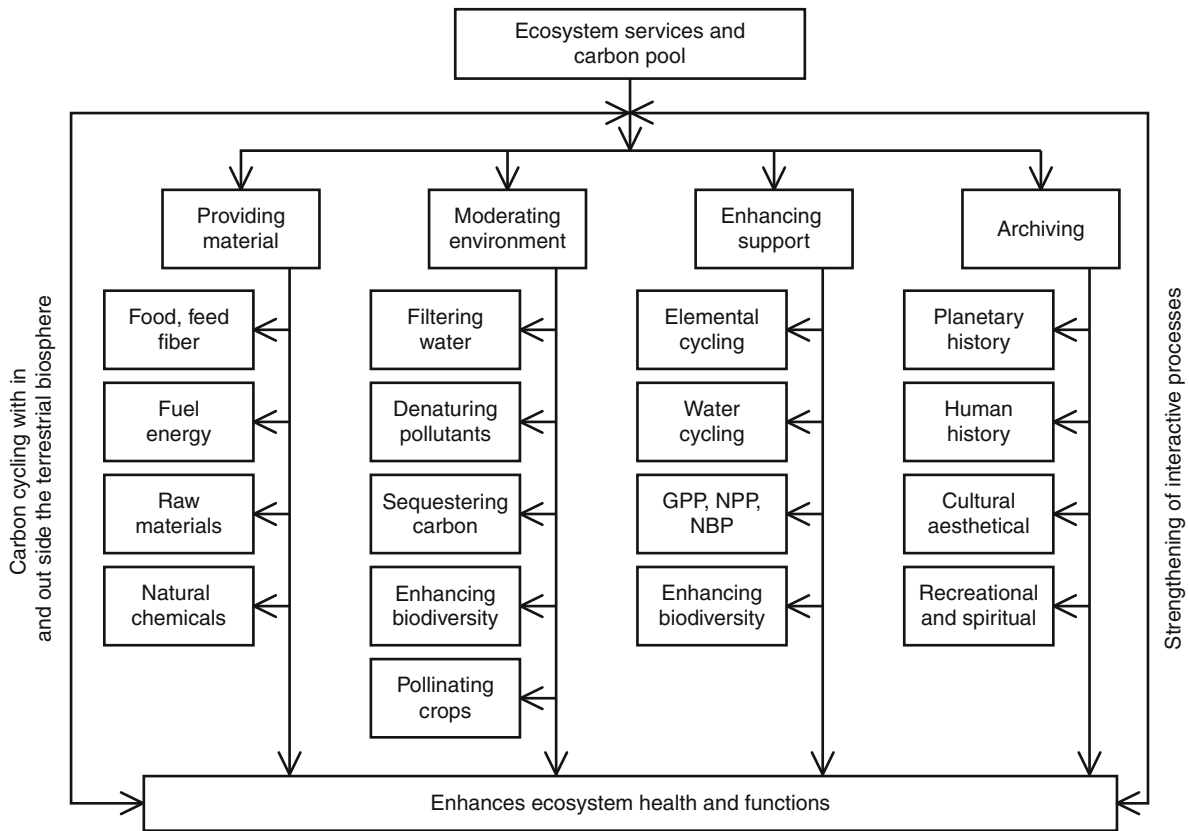
**Ecosystem Services and Soil Carbon Sequestration**

Sequestration of C in soils and biota generates and enhances numerous ecosystem services (Fig. 3). Important among these are: (1) providing materials of use to human (i.e., food), (2) moderating the environment (i.e., climate), (3) enhancing support, and (4) archiving human and planetary history [3, 4]. Soil and the ecosystem C pools are strong determinants of these services directly and indirectly. For example, quantity and quality of the SOC pool affect soil functions through (1) increasing soil aggregation and improving soil tilth, (2) increasing nutrient retention and availability, (3) moderating water retention and availability, (4) improving infiltration and reducing water runoff, (5) reducing soil erosion and nonpoint source pollution, (6) providing energy source and food to soil biota, (7) enhancing nutrient/elemental cycling, (8) accentuating use efficiency of input, (9) enhancing rhizospheric processes and micro-climatic environment, and (10) improving GPP and NPP.

**Carbon Sequestration and Agronomic Production**

It is because of numerous positive effects of organic carbon concentration in the root zone on soil quality that it is a strong determinant of the use efficiency of agronomic input and of crop growth and yield. Depending upon soil type and crop characteristics, there is a threshold value of 15–20 g/kg of SOC concentration in the root zone [5]. Crop growth and yield are strongly reduced when SOC concentration is below the threshold level (Fig. 4, [6, 7]). The yield response to SOC concentration in the root zone also depends on the management. The crop response (i.e., growth and yield) is generally stronger in agronomic practices based on low than high external inputs such as fertilizers, manure/compost, irrigation, etc. Indeed, the yield potential of elite varieties cannot be realized





**Agronomic Interactions with CO<sub>2</sub> Sequestration. Figure 3**

Ecosystem services enhanced by soil carbon sequestration (*GPP* gross primary productivity, *NPP* net primary productivity, *NBP* net biome productivity)

unless soil quality and the related agronomic interactions are optimized. Therefore, enhancing SOC concentration to above the threshold level is essential to improving agronomic yield in depleted/degraded soils of sub-Saharan Africa, South and Central Asia, and elsewhere in regions with low crop yields.

### Future Directions

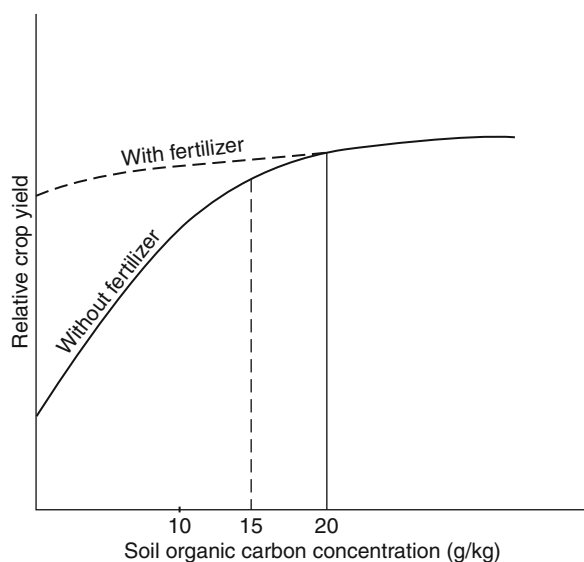
There is a strong need to understand the following:

- Soil processes and agronomic practices which create a positive C budget for diverse ecoregions
- Rate of C sequestration in relation to soil type, climate, and management
- Relationship between soil C pool/concentration and soil quality parameters, and agronomic yield
- Process, factors, and causes which enhance the mean residence time of C in soil

- Threshold value of C concentration for predominant soil type and crops
- Policy interventions to promote “carbon farming”

### Conclusions

Creating a positive budget of C (and N) is important to C sequestration in the soil. Agronomic practices to create a positive C budget are those which enhance the inputs of biomass-C. Important among these are conservation/no-till agriculture, mulch farming, cover cropping, integrated nutrient management, harvesting/recycling of water, complex farming systems, and perennial culture. Restoration of eroded/degraded soils and ecosystems is important to C sequestration in the soil. Energy-based inputs (i.e., fertilizers, pesticides, and tillage) have high hidden C costs. The strategy is to minimize losses of these



#### Agronomic Interactions with CO<sub>2</sub> Sequestration.

**Figure 4**

A generalized response curve of agronomic yield response of crops to concentration of soil organic matter in the root zone

inputs by erosion, leaching, volatilization, etc. Increase in soil organic C pool above the threshold level can enhance crop yield. Improvement in soil quality through increase in SOC pool is essential to increasing crop yields and agronomic production of soils in African, Asian, Caribbean, and the Andean regions.

#### Bibliography

1. Lal R (1999) Soil management and restoration for C sequestration to mitigate the greenhouse effect. *Prog Env Sci* 1:307–326
2. Lal R (2004) Carbon emissions from farm operations. *Environ Int* 30:981–990
3. Daily GC (1997) *Nature's services: societal dependence on natural ecosystems*. Island Press, Washington DC, 392 pp
4. MEA (2005) *Ecosystems and human wellbeing: synthesis, Millennium ecosystem assessment*. Island Press, Washington, DC, 155 pp
5. Aune JB, Lal R (1997) Agricultural limits in the tropics and critical limits of properties of Oxisols, Ultisols and Alfisols. *Trop Agric (Trinidad)* 74:96–103
6. Lal R (2010) Beyond Copenhagen: mitigating climate change and achieving food security through soil carbon sequestration. *Food Security* 2(2):169–177
7. Lal R (2010) Enhancing eco-efficiency in agroecosystems through soil C sequestration. *Crop Sci* 50:S120–S131

## Air Pollution Monitoring and Sustainability

ANDREW KNOX<sup>1</sup>, GREG J. EVANS<sup>1</sup>, COLIN J. LEE<sup>1</sup>,  
JEFFREY R. BROOK<sup>1,2</sup>

<sup>1</sup>Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada

<sup>2</sup>Air Quality Research Division, Atmospheric Science and Technology Directorate, Environment Canada, Toronto, ON, Canada

### Article Outline

Glossary

Definition of the Subject

Introduction

Air Pollutants

Currently Recognized Air Pollution Problems

Motivations Behind Monitoring Air Pollution

Current Air Pollution Monitoring Practices

Past and Ongoing Success in Air Pollution Mitigation

Outstanding Problems and Challenges

Current Air Pollution Monitoring and Sustainability

Future Directions: Monitoring to Promote Greater Sustainability

Conclusion

Bibliography

### Glossary

**Acid deposition** Settling on land or water of acidic air pollutants or air pollutants capable of becoming acidic once settled.

**Acid neutralizing capacity (ANC)** The ability of a medium (e.g., soil) to neutralize acidic pH, generally indicated by concentration of base cations.

**Adaptive capacity** The capacity to adjust to new circumstances such that successful functioning can continue.

**Aerodynamic diameter** The diameter of a spherical particle of unit density with a terminal velocity equal to the particle being considered.

**Aerosol** Liquid- or solid-phase particles suspended in a gas.

**Aerosol optical depth** Total extinction of light due to scattering and absorption by particles suspended in air between the point considered, the light's origin, and the point of observation.

**Air pollution monitoring** The use of scientific instruments and methods to assess the concentrations or rates of emission of air pollution in various environments.

**Air pollution policy** Legal guidelines governing allowable emissions of pollutants to air, allowable concentrations of pollutants in air, and air quality remediation efforts.

**Air quality monitoring** Observation of features of air that relate to the air's quality over some time period. Said features include concentrations of pollutants.

**Air quality** The degree to which air is an ideal supporter of life.

**Background level** The concentration of a substance or pollutant that remains regardless of short-term, local changes in emissions patterns. Background concentrations are not static, but changes in background concentration happen over the relatively long term.

**Best available technology (BAT)** A term relating to emissions standards set based on the ability of existing pollution abatement technology to mitigate the release of pollutants.

**Biogenic** Produced by biological organisms or systems.

**Biosphere** All life and ecosystems on Earth.

**Chlorofluorocarbons (CFCs)** A class of chlorine or fluorine containing organic compounds.

**Climate change** A long-term change in regional and global meteorological mean state or the variability of that mean state.

**Command and control** A strategy for limiting emissions by assigning specific emissions maximums to pollution sources or otherwise imposing pollution limiting rules and enforcing those rules.

**Concentration-response functions (CRFs)** The mathematical relationships between the environmental concentration of a pollutant and an effect of the pollutant. Effects, or responses, could include health effects such as premature mortality.

**Consolidated sustainability metrics** Summative statistics that are a function of a number of data on

different species of pollutants and atmospheric conditions. Consolidated metrics are based on a prior judgment of the relative importance of all the data in their calculation. They are designed to give an overall picture and convey a great deal of information in a single number or index.

**Critical load** The maximum amount of pollution an ecosystem can tolerate without suffering significant harmful effects, as far as can be understood given the present understanding.

**Critical natural capital** Natural capital that cannot be lost without irreparable terminal damage to ecosystems.

**Ecosystem** A system supporting life, including living and nonliving components and the interactions between its components.

**Enforcement** Action taken or penalties imposed to correct situations in which emissions or concentrations of pollutants are above the legal limit.

**Exposure** A time-integrated concentration metric reflecting the potential for intake of pollutants.

**Greenhouse gas** Gas which, when in the atmosphere, induces a climate forcing via its ability to absorb light and convert it to thermal energy.

**Hazardous air pollutants (HAPs)** Individual or families of compounds that can potentially induce adverse health impacts, including but not limited to cancer, reproductive effects, or adverse environmental effects.

**Holism** The view that systems cannot be understood in exclusion of other systems but must be understood in relation to the larger whole that affects all parts within itself.

**Mortality** A statistical measure quantifying the number or fraction of deaths among populations.

**Multipollutant monitoring** Monitoring with attention to the effects of pollution arising from the interaction of different species of pollutants.

**Natural capital** The natural systems that produce goods such as clean air and provide services such as life support to the biosphere.

**Ozone layer** A lower stratospheric region of more concentrated ozone, which absorbs much of the ultraviolet radiation from the sun.

**Particulate matter** Solid or liquid particles suspended in air. PM is not a term that distinguishes particles by chemical species, and individual PM

particles can contain a variety of species. Particulate Matter is often written as  $PM_{10}$  or  $PM_{2.5}$  with the number in subscript denoting the maximum aerodynamic diameter in micrometers of the particles being described. The subscripts (e.g., “2.5”) indicate that the PM considered is less than the subscript’s value in micrometers ( $\mu m$ ). Particles smaller than about 10  $\mu m$  are generally considered inhalable with smaller particles reaching deeper into the lung.

**Pedestrian-oriented development** Construction or reconstruction of the urban form designed to facilitate walking as a primary means of personal transportation.

**Point sources** Localized emitters of large amounts of pollution.

**Population exposure** A statistic accounting for a population’s distribution throughout the area it inhabits, the pollution concentration distribution over that area, and the changes in those distributions over time to represent the macro-level (population-level) exposure of the population to a pollutant or set of pollutants.

**Primary pollutants** Pollutants released from a source directly into the environment in the given form.

**Primary standards** US EPA Air pollution standards set to protect public health.

**Radiative balance** The net radiation to or from an object or system accounting for total incoming and outgoing radiation. A positive radiative balance will lead to the warming of the system considered, while a negative radiative balance will lead to cooling, assuming no non-radiative energy flows are at work.

**Resilience** The ability of a system to renew and reorganize itself after a shock or major socio-ecological shift.

**Risk-based policy** Policy based on the quantification and understanding of the probabilities of damages incurred. In risk-based policy, allowable pollutant concentrations are set with the knowledge and acceptance of the damage likely to be realized.

**Secondary pollutants** Pollutants not released directly from sources but formed in the environment via chemical reactions among primary pollutants facilitated by atmospheric conditions such as moisture, temperature, and sunlight.

**Secondary standards** US EPA Air pollution standards set to mitigate regional environmental problems.

**Socio-ecological justice** Fairness concerning both the social realm (e.g., fairness among social classes) and the ecological realm (e.g., distribution schemes, including ecosystems as stakeholders), and the interaction between the two realms.

**Source apportionment** Attributing the pollutants in a region or at a receptor site to their sources, for example, quantification of the portion of ambient  $NO_x$  originating from the transportation sector versus the industrial sector.

**Target load** The pollution burden aimed for in policy. The load deemed achievable once “other factors (such as ethics, scientific uncertainties, and social and economic effects) are balanced with environmental considerations.” [96]

**Transit-oriented development** Construction or reconstruction of the urban form designed to facilitate use of public transportation vehicles as a primary means of medium to long-distance personal transportation.

**Troposphere** The layer of the atmosphere closest to the Earth’s surface, where human exposure to air pollutants occurs.

**Urban form** The physical shape and design of cities, their structures, and their infrastructure.

## Definition of the Subject

Air pollution monitoring is the use of scientific and mathematical tools and methods to assess the concentrations or rates of emission of air pollution in various environments. Monitoring is usually performed at ground level in populated areas, where humans exposure tends to occur but can also include measurements at other altitudes, in other spheres of the atmosphere, or in unpopulated areas. Monitoring data is often an important input into air pollution research projects.

## Introduction

Air pollution monitoring is concerned with collecting evidence about the emissions, concentrations, and effects of air pollutants. There are a number of effects of air pollution. Those most commonly recognized are negative impacts on human health, ecosystems, and the economy. Current air pollution monitoring assesses

the properties of air pollution in the atmosphere by monitoring ambient air, monitoring emissions from pollution sources, remotely sensing air pollution indicators via satellite, and through the use of numerical models. Monitoring efforts allow for air quality warnings, tracking trends over time, air pollution research, and policy development, including pollutant concentration and emissions standards. While air pollution monitoring has led to many substantial successes in the reduction of impacts from air pollution, many air pollution problems are still outstanding.

Air pollution monitoring is an essential activity in the assessment of environmental quality and provides information without which acting to promote sustainability would be impossible. However, air pollution monitoring must be expanded in breadth and depth if societies are to progress toward sustainability. Monitoring must extend the information provided in service of more diverse goals. These goals should include facilitating urban design for air quality, socio-ecological justice, methods for acceleration of air pollution research, resilience, and preservation of natural capital. Work that crosses intellectual disciplines would also help, as would public involvement in monitoring. Immediate steps the air pollution monitoring community could take to improve its routine practice include the development of multipollutant impact assessments, greater consideration of consolidated metrics, more and better use of satellite-derived observations, routine use of available design methodologies in the design of monitoring networks, adhering to suggestions in existing literature to improve routine practice, and beginning to address some of the previously necessary oversimplifications involved in conventional monitoring. While the air pollution monitoring community has a historical record of promoting sustainability and should be applauded, a paradigm shift is needed if it is to address these expanded goals.

### Air Pollutants

Air pollution monitoring is concerned with air pollutants, substances which have negative effects on life. Some air pollutants are prevalent in most urban areas, and their effects are widespread. Other air pollutants tend to be less concentrated, or are only concentrated in specific areas, but have gained attention for being

especially problematic or acutely toxic. In the USA and in Canada, pollutants of this first type are classified as “criteria” pollutants. (In the USA, the label “criteria” is applied to common pollutants. The EPA regulates these pollutants by setting permissible levels based on human health or environmental considerations [1], whereas in Canada, pollutants fall into the “criteria” category as a result of sharing common sources, mostly notably, fossil-fuel combustion [2].) The second type of air pollutant, the type capable of causing problems despite relatively low average emissions rates, have been classified by the US EPA (US Environmental Protection Agency) as “Hazardous Air Pollutants” (HAPs) for the particular hazards they present. One hundred and eighty-nine HAPs are officially recognized by the US EPA [5]. Criteria pollutants and HAPs are discussed below, and classifications by different authorities are given in Table 1. Typical ranges for many non-HAP pollutants in the USA are given in Table 2. Table 3 summarizes the sources and effects of pollutants.

### What Are the Criteria Pollutants?

Criteria pollutants include carbon monoxide (CO), nitrogen oxides (NO<sub>x</sub>), and sulfur dioxide (SO<sub>2</sub>) [1]. Tropospheric ozone (O<sub>3</sub>) and Particulate Matter (PM) are also criteria pollutants. Most species of harmful Volatile Organic Compounds (VOCs) that are not acutely toxic enough to be specifically identified as HAPs are included in the criteria air contaminants in Canada [2]. Table 3 shows the common sources of the criteria pollutants mentioned and comments on some of their effects.

Emissions of CO are the result of the incomplete combustion of carbonaceous fuels. Any use of fossil fuels can result in emissions of CO, but vehicles are the dominant source [6]. In motor vehicles, catalytic converters are often used to mitigate CO emissions. CO's effects include human health impacts and contribution to the formation of tropospheric ozone [7].

NO<sub>x</sub> is the sum of NO and NO<sub>2</sub>. Atmospheric nitrogen (N<sub>2</sub>) is oxidized during combustion to form NO<sub>x</sub>. NO<sub>x</sub> is similar to CO in that it can lead to health effects and the production of tropospheric ozone, its main source in many locations is emissions from vehicles, and catalytic converters are used to minimize these



**Air Pollution Monitoring and Sustainability. Table 1** Pollutants and their classification by different systems. Pollutants are described in detail below. The more acutely hazardous pollutants given here (those classified as HAPs by the US system, for example) are only examples and not an extensive representation of the full range of hazardous pollutants. PM<sub>10</sub> and PM<sub>2.5</sub> are classified here as CACs by the Canadian classification and Criteria Pollutants by the US EPA classification, as Environment Canada and the US EPA classify “particulate matter” in general as CACs and Criteria Pollutants, respectively [2]. Certain species of dioxins and furans are classified as HAPs by the EPA, but not all are covered. CAC stands for criteria air contaminant, AQSP stands for air quality strategy pollutant, POP stands for persistent organic pollutant [3], HAP stands for hazardous air pollutant

| Pollutant          | Canadian classification [2] | UK classification [4] | US classification [5] |
|--------------------|-----------------------------|-----------------------|-----------------------|
| CO                 | CAC                         | AQSP                  | Criteria pollutant    |
| NO <sub>x</sub>    | CAC                         | AQSP                  | Criteria pollutant    |
| SO <sub>2</sub>    | CAC                         | AQSP                  | Criteria pollutant    |
| NH <sub>3</sub>    | CAC                         |                       |                       |
| O <sub>3</sub>     | CAC                         | AQSP                  | Criteria pollutant    |
| VOC                | CAC                         |                       |                       |
| PM <sub>10</sub>   | CAC                         | AQSP                  | Criteria pollutant    |
| PM <sub>2.5</sub>  | CAC                         | AQSP                  | Criteria pollutant    |
| Dioxins and Furans | POP                         |                       | HAP                   |
| 1,3-butadiene      |                             | AQSP                  |                       |
| Benzene            |                             | AQSP                  |                       |
| Lead               |                             | AQSP                  | Criteria pollutant    |
| Arsenic            |                             |                       | HAP                   |
| Cadmium            |                             |                       | HAP                   |
| Mercury            |                             |                       | HAP                   |
| Nickel             |                             |                       | HAP                   |

emissions. Other high-temperature combustion processes also contribute. Emissions of NO<sub>x</sub> can indirectly lead to acid deposition (or “acid rain”).

Combustion of fossil fuels containing sulfur leads to emissions of SO<sub>2</sub>. Refining and production of fuels and minerals (e.g., lead smelting, aluminum refining) also emits substantial amounts of SO<sub>2</sub>. SO<sub>2</sub> negatively impacts human health, contributes to acid deposition, and can promote aerosol formation and growth. Sulfate formed from SO<sub>2</sub> is a major contributor to aerosol mass in some regions. Aerosol produced through historic emissions of SO<sub>2</sub> is believed to have caused a negative climate forcing that suppressed temperature rise in the second half of the last century. Using fuels that contain only trace amounts of sulfur, such as natural gas, or that contain less sulfur, such as low-

sulfur coal, results in less emission of SO<sub>2</sub>. There are also downstream options such as flue-gas desulfurization [6].

Tropospheric ozone (O<sub>3</sub>) is a major irritant to human health and vegetation. O<sub>3</sub> is a secondary pollutant, meaning that O<sub>3</sub> is rarely directly emitted from point sources but is formed via chemical reactions involving other air pollutants such as NO<sub>x</sub> and VOCs.

There is a multitude of volatile organic carbon (VOC) species both biogenic and anthropogenic. Many have been identified as particular hazards, and they are discussed in the section below on HAPs. Some anthropogenic VOCs have been associated with causing cancer in humans [8]. Anthropogenic sources include vehicle emissions, petroleum product manufacture and use, biomass burning, landfill gas, sewage treatment [9],

**Air Pollution Monitoring and Sustainability. Table 2**

Typical concentrations of pollutants in the USA [12]

| Pollutant                           | US average range                              | US typical peak <sup>c</sup> |
|-------------------------------------|-----------------------------------------------|------------------------------|
| CO <sup>a</sup>                     | 0.1–5 ppm                                     | 20 ppm                       |
| NO <sub>2</sub> <sup>a</sup>        | 0.5–50 ppb                                    | 200 ppb                      |
| NO <sup>a</sup>                     | 0–100 ppb                                     | 200 ppb                      |
| HNO <sub>3</sub> <sup>a</sup>       | 0–5 ppb                                       | 10 ppb                       |
| SO <sub>2</sub> <sup>a</sup>        | 0.1–50 ppb                                    | 150 ppb                      |
| PM <sub>2.5</sub> Lead <sup>b</sup> | 0.5–5 ng/m <sup>3</sup>                       | 150 ng/m <sup>3</sup>        |
| O <sub>3</sub> <sup>a</sup>         | 0–125 ppb                                     | 200 ppb                      |
| (VOC) <sup>b</sup>                  | 20–100 µg/m <sup>3</sup>                      | 250 µg/m <sup>3</sup>        |
| PM <sub>10</sub> <sup>b</sup>       | 10–100 µg/m <sup>3</sup>                      | 300 µg/m <sup>3</sup>        |
| PM <sub>2.5</sub> <sup>b</sup>      | 5–50 µg/m <sup>3</sup><br>(mean = 13.4 ± 5.6) | 100 µg/m <sup>3</sup>        |
| UFP <sup>a</sup>                    | 1,000–20,000/cm <sup>3</sup>                  | 100,000/cm <sup>3</sup>      |

<sup>a</sup>Typical hourly average concentrations reached in US cities<sup>b</sup>Typical 24 h average concentrations<sup>c</sup>Generally not in concentrated plumes or locations of direct source emission impact

and off-gassing from building materials. Biogenic VOCs can be the result of the natural metabolic processes of life forms [10] and natural decomposition. VOCs play a crucial role in the formation of tropospheric ozone [11].

PM (particulate matter) is a term that covers a wide variety of liquid and solid particles suspended in the atmosphere. PM is classified by size (according to aerodynamic diameter) and various chemical species of PM exist, even within single particles. Size classifications are typically given as PM<sub>10</sub> (particles with aerodynamic diameters less than 10 µm), PM<sub>2.5</sub> (particles with aerodynamic diameters less than 2.5 µm), and Ultrafine Particles (UFP, particles with aerodynamic diameters less than 0.1 µm). Particulate species include sulfate (SO<sub>4</sub><sup>2-</sup> and containing particles), nitrate (NO<sub>3</sub><sup>-</sup>), ammonium (NH<sub>4</sub><sup>+</sup>), Organic Matter (OM), and Black Carbon (BC). Long-term exposure to PM<sub>2.5</sub> is one of the major immediate human health risks from air pollution [13]. Specific health effects are varied but include cardiovascular and pulmonary complications. Other effects of PM include the ability to alter the radiative balance of the Earth [14].

### What Are the HAPs?

Specific VOCs are very hazardous to human health. These include benzene, PAHs (polycyclic aromatic hydrocarbons), 1,3-butadiene, ethane, propene, and some aldehydes [6]. Hazardous VOCs share sources with the bulk of VOCs.

Dioxins and furans are chlorine containing chemicals with similar chemical structural elements [15]. Their health effects can be extreme and of a wide variety, including cancer and reproductive abnormalities [15]. Dioxins and furans tend to accumulate in soils and sediments, but their sources include sources of air pollution such as municipal waste incinerators [15] and pyrotechnics [16]. They are bio-accumulative [17], meaning that they remain within the bodies of plants and animals that absorb them.

Heavy metals such as arsenic and nickel are not very prevalent in the atmosphere, other than near their industrial sources, but even at low concentrations they can produce serious health effects [6]. These metalloids, metals and their compounds are classified as HAPs. Some lighter metals are also classified as HAPs such as beryllium and cobalt [5].

Lead (Pb) is a HAP [5] worthy of note for its relative abundance. Lead emissions are the source of a wide range of human health problems [18, 19]. While lead as a fuel additive has largely been phased out in the Western world, there are many regions where it is still in use and continues to exist in high concentrations in the atmosphere [6]. Lead is classified both as a HAP [5] and a criteria pollutant in the USA [1].

Mercury (Hg) is a HAP [5] with a number of industrial applications, many still in use today [20] and is a component of compact fluorescent light bulbs. Coal combustion-based electricity generation is one of mercury's major atmospheric emission sources. Mercury is toxic to humans and, once in the environment, can be methylated, whereupon it can easily bind to proteins and enter the food chain [20].

### Currently Recognized Air Pollution Problems

Before resources can be committed to monitoring air pollution, the nature and extent of the problems air pollution creates must be understood. Fenger [6] cites three types of air pollution problems. These are local problems such as health effects and damage to goods

**Air Pollution Monitoring and Sustainability. Table 3** Criteria pollutants, HAPs, their common sources, and effects. (VOCs are considered CACs in Canada but are not on EPA list of Criteria Pollutants)

| Substance                                             | Sources                                                                                                                                                  | Effects                                                                                            |
|-------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|
| Criteria pollutants                                   |                                                                                                                                                          |                                                                                                    |
| Carbon Monoxide (CO)                                  | Combustion of fossil fuels, especially in motor vehicles                                                                                                 | Health effects, smog (Ozone) formation                                                             |
| Nitrogen Oxides (NO <sub>2</sub> )                    | Combustion of fossil fuels                                                                                                                               | Health effects, smog (Ozone and PM <sub>2.5</sub> ) formation, and acid deposition                 |
| Sulfur Dioxide (SO <sub>2</sub> )                     | Combustion of sulfur containing fossil fuels and industrial activities (e.g., fuel refining, metal production)                                           | Health effects, acid deposition, smog (PM <sub>2.5</sub> ) formation, changes in radiative balance |
| Tropospheric Ozone (O <sub>3</sub> )                  | Chemical reactions involving other pollutants in the atmosphere                                                                                          | Health effects, damage to vegetation, damage to materials                                          |
| Volatile organic compounds (VOCs)                     | Biogenic, vehicle emissions, petroleum products, biomass combustions, landfill gas, sewage treatment, off-gassing of building materials                  | Health effects, smog (Ozone and PM <sub>2.5</sub> ) formation                                      |
| Particulate Matter (PM)                               | Varies by species, includes combustion emissions (especially of biomass in wildfires and for fuel), secondary chemistry involving gaseous air pollutants | Health effects, visibility, changes in radiative balance                                           |
| HAPs                                                  |                                                                                                                                                          |                                                                                                    |
| Hazardous VOCs                                        | Varies by VOC species                                                                                                                                    | Health effects, including cancer, smog (Ozone and PM <sub>2.5</sub> ) formation                    |
| Hazardous Metals (excluding lead) and their compounds | Industrial sources                                                                                                                                       | Health Effects, soil contamination                                                                 |
| Lead                                                  | Fuels with lead additives, industrial sources                                                                                                            | Health Effects, soil contamination                                                                 |

and capital, regional problems such as acidification, and global problems such as climate change. As air pollution problems scale up in physical area (from regional to global, for example), their effects tend to scale up in terms of time. For example, an air quality episode may aggravate asthma in a population over the course of the episode (hours to days), whereas acidification will impact the environment for decades and climate change will have impacts potentially lasting millennia [21]. The problems receiving the most government recognition are the local problems. Regional problems are recognized by most governments, while some global problems are only recently appearing on the air pollution agenda of some authorities (for example, the US EPA only gained powers to regulate CO<sub>2</sub> emissions as of December, 2009).

Nowhere is the hierarchy of air pollution problems more apparent than in the US EPA's division between primary and secondary air quality standards. Primary standards are set to protect public health, while secondary standards are set to mitigate regional problems, such as acidification, and biosphere damage [1]. There are no tertiary standards set to mitigate global problems. However, in Europe, the "EU Focus on Clean Air" [22] lists climate change and the destruction of the ozone layer among air pollution issues, placing global problems alongside childhood asthma and acidification of forests.

In addition to human health and regional problems, the economic problems caused by air pollution are recognized by governments. For example, Environment Canada recognizes that illness and mortality resulting from air pollution can cause workers to miss work and

thereby have an adverse affect on the productivity of industry [23]. Economic problems arising from air pollution are discussed below along with other problems currently recognized by government agencies.

## Health

Human health is often the primary concern related to air pollution, and a great deal of work has been done to connect various air pollutants to health risks. Table 4 identifies some of the health problems associated with the different air pollutants. The references given are examples of studies that have made connections between pollutants and health problems. Thus, Table 4 should not be interpreted as an exhaustive and critical review of the current state of knowledge regarding air

pollutant health effects. In some cases, many more studies have confirmed these connections. However, in many cases, the underlying mechanisms that would conclusively demonstrate a cause and effect relationship have yet to be identified. At the present time, PM<sub>2.5</sub> is considered to be one of the pollutants of greatest concern [12, 32, 42], but it is made up of thousands of chemicals that can each have different types of effects. Ultrafine particles are a subsize of PM<sub>2.5</sub>, and they warrant further study to better understand the range of impacts on health they may cause.

The health effects of air pollution are usually more pronounced in urban populations [6]. Urban settings tend to host more combustion (of transportation and heating fuels, for example) per area, leading to more emission of air pollution per area than in rural settings.

**Air Pollution Monitoring and Sustainability. Table 4** Air pollutants and their health effects. COPD stands for chronic obstructive pulmonary disease. This table

| Pollutant                         | Associated health problems                                                                                                                                                                                                                                                  | Reference       |
|-----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| NO <sub>2</sub>                   | COPD, allergic response, increase in respiratory illness, increased airway responsiveness to stimuli, lung inflammation, decreased pulmonary function, immunological changes, eye irritation                                                                                | [19, 24, 25]    |
| SO <sub>2</sub>                   | COPD, respiratory symptoms                                                                                                                                                                                                                                                  | [19, 24]        |
| CO                                | COPD, Congestive heart failure, decreased time to onset of angina, behavioral effects, developmental effects                                                                                                                                                                | [19, 24]        |
| O <sub>3</sub>                    | Lung inflammation, COPD, aggravation of asthma, eye and nose irritation, pneumonia, increased airway responsiveness to stimuli, centroacinar fibrosis, immunological changes, excess mortality                                                                              | [19, 24, 26–28] |
| PM <sub>10</sub>                  | Exacerbation of cardiac disease, lung cancer, allergic response, infant mortality                                                                                                                                                                                           | [25, 29–32]     |
| PM <sub>2.5</sub>                 | Acute and chronic cardiovascular effects, increased incidence of allergies, excess mortality, lung cancer                                                                                                                                                                   | [12, 32–34]     |
| Ultrafine PM                      | Cardiorespiratory mortality, pulmonary stress, mitochondrial and DNA damage, potentially other effects as yet proven                                                                                                                                                        | [35–39]         |
| Manganese (Mn)                    | Parkinson's Disease                                                                                                                                                                                                                                                         | [40]            |
| Mercury (Hg)                      | Nervous system damage                                                                                                                                                                                                                                                       | [41]            |
| Lead (Pb)                         | Reproductive problems, behavioral deficiencies in offspring, damage to lungs, nervous system damage, lowered IQ, hypertension, coronary heart disease, reproductive effects in women                                                                                        | [18, 19]        |
| Volatile organic compounds (VOC)s | Cancer                                                                                                                                                                                                                                                                      | [8]             |
| Dioxins and furans                | Cancer, deficiencies in the immune system, reproductive and developmental abnormalities, nervous system damage, endocrine disruption, pulmonary problems including bronchitis, altered testosterone levels, nausea, loss of appetite, rashes, liver damage, and many others | [15]            |

The higher population densities of urban settings results in greater air pollution population exposure. With over half of all people currently living in cities and the urban population expected to grow to 70% of all people by 2050 [42], the health impacts of air pollution will continue to be a challenge. Long-range transportation of air pollution and local point sources can make the health effects of air pollution a significant risk in rural settings as well.

A cohort study of six US cities started in 1974 and followed up through the present time, the “Harvard Six Cities Study,” found that the average life expectancy in the dirtiest city studied (corrected for individual risk factors) was 2 years lower than in the least dirty city [43]. This 2 year gain in average life expectancy is equivalent to what would be expected if all forms of cancer in the USA were eliminated [43]. Indeed, as air quality has improved across the USA in recent decades, life expectancy has risen [44]. Thus it can be seen that the health effects of air pollution pose a substantial problem while mitigation can offer substantial benefits.

## Regional Problems

Regional air pollution problems impact large areas and substantially alter ecologies. Regional problems that have been recognized by governments include acid deposition, damage to wildlife and ecosystems, and reduced visibility. In some areas of the globe, many of these problems have been substantially mitigated by changes in practice.

**Acid Deposition** Acid deposition is the settling of acidic air pollutants on land or water. The pollutants contributing most significantly to acid deposition are oxidized sulfur and nitrogen compounds. When “wet deposited” (deposited on land by precipitation after being taken up in cloud droplets) sulfur is usually in the form of  $\text{SO}_4^{2-}$  due to oxidation of  $\text{SO}_2$  during atmospheric transport. When “dry deposited” (directly deposited from air onto Earth’s surface), it is usually in the forms of  $\text{SO}_2$  and particulate  $\text{SO}_4^{2-}$  [45].  $\text{NO}_x$  reacts with water to form acids either while airborne or after dry deposition.

On a regional scale, farther downwind from emission sources, wet deposition is the major acid deposition mechanism and has earned acid deposition the

popular name “acid rain,” though dry deposition makes substantial contributions to ecosystem acidification [45]. Acid deposition reduces ecosystem Acid Neutralizing Capacity (ANC) [46], making future influxes of acids more difficult to cope with. It also has deleterious effects on soils, vegetation, and aquatic ecosystems.

Leaching of soil nutrients is a major impact of acid deposition. Acids deposited in soil react with the soil’s base cations (largely  $\text{Ca}^{2+}$  but also  $\text{Mg}^{2+}$  and  $\text{K}^+$ ) and increase their mobility in soils, causing them to leach out of soils and making them unavailable for plant uptake [46]. Base cations can be the limiting factor in forest development and their loss has been identified as the source of decline in many forests [47]. Increased acidity in soils also negatively impacts the microbial community in the soil, on which plant life depends, notably decreasing biological nitrogen fixation rates [48]. Acid deposition stresses plant root systems and leaves [49] and can leach nutrients (Ca, Mg, and K) from leaves just as it does from soils [50]. Acid and/or nitrogen deposition have also been seen to increase risk of winter injury in Red Spruce [51]. Acid deposition lowers the pH of lakes and streams, negatively affecting organisms at all levels of the aquatic biosphere in terms of production and growth in lakes with low natural ANC [52]. Acute mortality and reproductive failure has been a result among fish and some species have even been lost entirely [52].

Governments in North America and Europe have addressed acid deposition by limiting the use of sulfur-rich fuels and introducing technologies to reduce emissions of chemicals that eventually result in acid deposition. Acid deposition has been substantially reduced as a result of these strategies, though many of its effects persist. China has not yet addressed acid rain to the extent that North American and European governments have, and today almost 30% of China’s area suffers from acid rain [53]. Given the transboundary nature of acid deposition, it is likely that much more area is threatened.

**Damage to the Biosphere** In addition to the damages to forests and wildlife caused by acid deposition, there are other damages caused by other forms of air pollution. Grantz et al. [54] provides Guderian’s categories of effects of air pollutants on ecosystems:



- Accumulation of pollutants in plants and other ecosystem components (e.g., soil, water)
- Damage to humans as a result of pollutant accumulation
- Shifts in competition causing changes in species diversity
- Disruption of biogeochemical cycles
- Disruption of stability and reduction in self-regulation capabilities
- Breakdown of stands and associations
- Expansion of denuded zones

Many of these types of impacts have been observed and linked to particular pollutants. For example, mercury has been shown to be toxic to microorganisms, plants, and animals [55], and toxicity of other heavy metals such as lead has also been confirmed [18]. By binding to proteins, mercury can also accumulate in organisms and ecosystems [20].

Ground-level ozone ( $O_3$ ) is a known hazard, leading to reduced production of fine roots in some trees [56], reduced tree growth [57], and damage to a plant's photosynthetic systems [58]. The reduced activity of plant life as a result of ozone will reduce an ecosystem's ability to store carbon and thereby exacerbates climate change [59]. Damage and reduced yields of plants will reduce the availability of food for wild animals, stressing those populations as well.

### Visibility

Air Pollution (mostly particles) can reduce visibility. Severe visibility problems arising from poor air quality have recently been experienced in some Chinese cities [60]. Human enjoyment of the environment is impeded by poor visibility and in extreme cases poor visibility poses a safety hazard in aviation.

### Economy

Restricted activity and lost work days due to illness incurred by air pollution are included in the costs of poor air quality [19]. These are costs both to the individual suffering the illness (in terms of income) and to industry (in terms of lost productivity). Time lost by caregivers, for example, the parents that miss work to care for an ill child, also results in financial costs [61]. Healthcare costs, including the fees of healthcare

professionals and operating costs, are also a problem. Pain and suffering has a cost that can be estimated by the price an individual is willing to pay to avoid experiencing it [61]. Similarly, the amount people are willing to pay to avoid risk of premature death can be used to estimate the cost of mortality [62].

Loss of detail in carvings, discoloration of materials, failure of protective coatings, and structural failure are some of the outcomes of material damage from air pollution [63]. These outcomes are the result of acid deposition on materials and oxidation from pollutants like ozone. Delucchi's [64] review of cost estimates in the literature yielded an estimate of \$0.1–3 billion in material damage annually in the USA from ozone.  $SO_x$  and PM are estimated to do at least as much damage as ozone [64]. Mayerhofer et al. [63] estimated that the damage from 1995 levels of air pollution cost Europe 2.3–5.3 trillion European Currency Units (ECU, 1995 value) in that year. Calcareous rock (marble and limestone) is vulnerable to air pollution [65], and many historical monuments and structures are made of such materials.

Ozone can enter plant leaves and damage enzymes critical to photosynthesis, reducing the metabolism of plants and limiting plant growth [66, 67]. Ozone, other oxidants,  $SO_x$ , and  $NO_x$  are thought to be responsible for 90% of damage to plants from air pollution [67]. Damage to crops results in less revenue for agricultural producers. Murphy et al. [66] estimate that in 1990, anthropogenic ozone caused \$2.8–5.8 billion in damage to 8 major crops in the USA (corn, soybeans, wheat, alfalfa hay, cotton, grain sorghum, rice, barley). Barley and soybeans are particularly sensitive to damage from  $SO_2$  [66]. Subsidies to these crops in the USA are on the order of billions [68], meaning that American taxpayers are providing substantial economic recompense for air pollution's damage to crops.

There are a number of services of economic value delivered by ecological systems, including dilution and treatment of wastes, employment, sequestration of carbon, provision of oxygen, and production of goods such as timber, potable water, and irrigation water [54]. The US EPA, 1997 [19] adds pesticides and drugs to the list of products derived from ecosystems as well as recreation, reduced runoff, and reduced erosion to the list of services of economic value. Efforts have been made to estimate some of the economic effects of air pollution's impacts on ecosystems such as wetlands, forests, and

aquatic ecosystems, but realistic estimates are difficult to reach [19]. Both the data and the tools to make complete estimates are lacking [19], as is the discourse to address some forms of ecosystem damage. The effects of local wilderness areas on property value are easier to estimate and work has been done in this area [69]. Such estimates can quantify some portion of the costs of the expansion of denuded zones and the limitations to ecosystem development discussed above. However, property values are just one of many economic impacts associated with pollution, leaving complete costs figures out of reach.

An evaluation of the cost-effectiveness of the Clean Air Act (USA) over the 1970–1990 period showed that the cost of compliance with the act was \$523 billion and gave the central estimate of the economic benefits as \$22.2 trillion [19]. This estimate of benefits does not include complete evaluations of ecosystem benefits. The 1990 amendments to the Clean Air Act created market-based mechanisms for trading of SO<sub>2</sub> in an effort to reduce emissions and emissions were reduced by 40% between 1990 and 2007. Over the same period, 25 member countries of the EU reduced SO<sub>2</sub> emissions by 70% using conventional, non-market-based legislation [70].

## Global Problems

Air pollution and its effects do not respect national boundaries. Regardless of the location of the point of emission, some problems arising from air pollution affect multiple countries. Some, like the destruction of the ozone layer and climate change, are a problem for all life on Earth.

**Ozone Layer Destruction** While tropospheric ozone can be harmful, the stratospheric ozone layer prevents much of the sun's ultraviolet radiation from reaching the Earth's surface. Anthropogenic chlorofluorocarbons (CFCs) emitted to the atmosphere can act to convert stratospheric ozone into oxygen through a series of reactions. Once the reactions are complete, the destructive capacity of the CFCs is left intact. The reaction series is able to begin again, destroying more ozone through a catalytic process [71]. In the mid-1980s, a large hole in the ozone layer above the Antarctic was documented, as was the general depletion of

stratospheric ozone. The damage to the ozone layer led to the Montreal Protocol, which legislated reductions in the production of ozone-depleting substances such as CFCs [72]. While existing ozone-depleting substances will, over decades, be removed from the atmosphere [71], other factors affecting stratospheric ozone chemistry, many of which are anthropogenic, make for some uncertainty about whether the ozone layer will ever return to the pre-1980s levels [72].

**Climate Change** Air pollutants such as methane and nitrous oxide are greenhouse gases and, as such, act to increase the Earth's trapping of electromagnetic radiation, leading to the increase in global average temperatures often referred to as global warming or climate change. Some Greenhouse Gases (GHGs) are affected by air pollution through chemical reactions [73], and GHGs such as CO<sub>2</sub> are co-emitted with primary pollutants, especially during fossil fuel combustion. Ozone in particular affects the lifetime of GHGs [73]. Other effects of climate change include such global problems as rising sea levels, likely extinction of 15–30% of species (with a 2°C increase in global average temperatures), reduced agricultural productivity, and fresh-water shortages [74].

## Summary of Air Pollution Problems

The motivating factors behind air pollution monitoring include understanding air pollution's impacts on human health, regional problems including economic costs, and global problems. Effects on health and the global economy are far from negligible, and the urgency of regional and global environmental problems is apparent. None of these problems at their current magnitude are compatible with sustainability. If the Earth's growing civilization is to be sustainable, understanding air pollution in an effort to minimize its direct and indirect impacts must be a priority.

## Motivations Behind Monitoring Air Pollution

Immediate benefits of air pollution monitoring include knowledge of when and where standards are violated and the ability to communicate health risks. Communication of health risk can be accompanied by warnings that can guide behavior of individuals, populations, and industry and reduce human exposure to air pollution.

While minimizing health problems and ensuring that standards are adhered to are important, there are deeper motivations for the monitoring of air pollution. Establishing and quantifying connections between air pollution and the related impacts depends on measurements such as those made through monitoring and related air pollution research. For example, epidemiological studies of associations between air pollution and related hospital admissions could not occur without air pollution monitoring. By proving connections between air pollution and problems, monitoring and the related research influence air pollution policy. It is anticipated, but not guaranteed, that continued improvement of policy will drive improved mitigation efforts and hence avoidance of the problems associated with air pollution.

**Air Quality Warnings**

Monitoring of air pollution yields data on present air quality conditions that can be used to provide health warnings. For example, Canada’s Air Quality Health Index (AQHI) combines data on ozone, PM, and NO<sub>2</sub> to produce an AQHI value ranging from 1 to 10+. Advice associated with the AQHI is given in Table 5 and ranges from “Enjoy” usual activity to “Avoid” strenuous outdoor activity.

Ideally, this health advice, if followed, would reduce the impact of air pollution by reducing the rate of

exposure. Air Quality Indices designed to inform the public and potentially alter behavior exist in the USA, China, Europe, and other areas of the globe [13].

**Research**

Research serves the function of improving the understanding of air pollution, how it behaves, its sources, and its connection to the problems it causes. It can also lead to advancement of monitoring techniques. In an effort to classify research questions typically asked in the field of air pollution research, a survey of research questions from 40 articles pertaining to air pollution published in early 2010 in peer reviewed journals was performed. Research questions appearing in this snapshot of the literature tend to be related to the following broad questions:

1. What is the nature of a given pollutant or set of pollutants (in terms of pollutant properties, distribution, and transport patterns in the atmosphere)?
2. Can we improve our understanding of the factors influencing air quality (e.g., meteorological factors)?
3. What chemical or physical mechanisms explain some yet unexplained observations?
4. At what rate and by what process is a given pollutant rendered benign in the environment?
5. What are the health effects of a specific pollutant or set of pollutants?

**Air Pollution Monitoring and Sustainability. Table 5** Canadian air quality health index (AQHI) and related health advice [75]

| Air quality health index | Health advice                                                                                                    |                                                                                                                                          |
|--------------------------|------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
|                          | At risk populations                                                                                              | General population                                                                                                                       |
| 1–3                      | <i>Enjoy</i> your usual outdoor activities                                                                       | <i>Ideal</i> air quality for outdoor activities                                                                                          |
| 4–6                      | <i>Consider reducing</i> or rescheduling strenuous activities outdoors if you are experiencing symptoms          | <i>No need to modify</i> your usual outdoor activities unless you experience symptoms such as coughing and throat irritation             |
| 7–10                     | <i>Reduce</i> or reschedule strenuous activities outdoors. Children and the elderly should also take it easy     | <i>Consider reducing</i> or rescheduling strenuous activities outdoors if you experience symptoms such as coughing and throat irritation |
| 10+                      | <i>Avoid</i> strenuous activities outdoors. Children and the elderly should also avoid outdoor physical exertion | <i>Reduce</i> or reschedule strenuous activities outdoors, especially if you experience symptoms such as coughing and throat irritation  |

6. What is a population's or an individual's exposure to a specific pollutant or set of pollutants?
7. What are the effects of a specific pollutant or set of pollutants on outcomes other than health?
8. What is a given practice's effect on air pollution levels (e.g., how does highway driving affect air pollution levels)?
9. How are levels of air pollutants changing (are they decreasing as expected given current policies or are they rising in particular regions)?
10. How does a specific instrument, method, or type of analysis perform?

Investigations of these questions inform the air quality monitoring community, are informed by the data collected by said community, and guide interpretations of data. They provide information on the nature of air pollution, the connections between air pollution and problems, and the effects of human behavior on air pollution. They also allow progress in the air pollution monitoring field.

### Questions Regarding the Nature of Air Pollution

Research questions of types 1–4 are important to the development of the understanding of the nature of air pollution and specific pollutants. Fundamental questions such as these aid in design of mitigation strategies, monitoring strategies and technologies, and can inform policy. They can also suggest reasons why pollution is a problem and which kinds of problems can be expected from particular pollutants.

Questions of type 4 are directly relevant to sustainability. *A sustainable system would not involve emissions rates higher than the rate at which pollutants are rendered benign. Answers to questions of type 4 should inform allowable rates of emissions.*

### Questions Connecting Air Pollution to Problems

Question types 5–7 help to establish connections between air pollution and the problems that make air pollution a subject of concern and make monitoring a necessary mitigation strategy. Without questions such as these, other types of air pollution questions lose much of their practical application. These questions are central in “risk-based” air pollution

management strategies, which frame acceptable levels of pollution in terms of the degree to which their effects are deemed acceptable.

Examination of these questions shows how established air pollution objectives affect the research agenda. As has been discussed, governments tend to emphasize the impacts of air pollution on health, some selected environmental issues, and the economy. As a result, research questions of types 5–7 more frequently attempt to make connections to these issues as opposed to issues given less attention such as the effects of air pollution on overall quality of life or the health of other species. (Climate change is a notable exception, however, and does receive much attention).

Question 7 covers a broad area of research and is often split into several questions investigating the effects of air pollution on climate change, acidification, ozone depletion, economic damages, and other unintended consequences of air pollution. *If the focus of air pollution monitoring is to shift toward sustainability, these types of questions will join health-related questions as questions of central importance.*

Questions connecting air pollution to impacts require continual study. Previously unrecognized or underappreciated issues may be found to be of great importance, thus warranting a change in monitoring. Furthermore, relationships between air pollution and its impacts can change over time, meaning that the nature of established relationships requires continual monitoring. Weatherhead and Anderson's 2006 [72] study is an example of the understanding of a relationship changing as a result of continued questioning. The assumption was that the ozone layer would regain its former abilities as CFC concentrations declined. However, other anthropogenic factors are playing a role in reducing the rate of recovery, despite a decline in CFCs [72]. Airborne PM is another example. Monitoring first dictated that total suspended particulate (TSP) be measured. However, monitoring evolved with increased understanding of health effects and the nature of PM. This increased understanding, itself facilitated by monitoring, recommended that attention be paid to specific size fractions of PM, first PM<sub>10</sub>, and later PM<sub>2.5</sub>. If connections to problems are not continually reviewed, monitoring can lose its relevance to its objectives.

### Research Questions and the Development of Air Pollution Monitoring

Questions about the nature of air pollution can aid development of new monitoring technologies and deployment strategies. For example, Miffre, et al. [84] reveals information about the differences between air pollution concentrations at ground level and at 50 m above ground. The findings show when it is valid to assume that ground-level monitoring is sufficient in estimating concentrations at higher elevations and when it is not. This study and studies asking similar questions directly inform how much information can be taken from a given monitoring practice and when more monitoring is necessary, increasing the potential for optimally efficient monitoring. Questions connecting air pollution to problems can guide monitoring so that it is more likely to take place at the point in a pollutants life-cycle that is most directly related to the problems of concern. If monitoring can take place, where it is most relevant to problems arising from air pollution, it will be more efficient and will yield the most relevant and useful results.

Questions of type 8 can direct monitoring to the areas where there is likely to be air pollution by developing an understanding of which practices are likely to create the most pollution and which are relatively benign, requiring less monitoring. Questions of type 9 bring about understanding of how changes in practice impact air pollution concentrations and whether current policies are effective. Type 10 questions can lead to development of new technologies and strategies for monitoring, as well as improvements to existing techniques.

### Current Air Pollution Monitoring Practices

Common pollutants (Criteria pollutants, CACs) are routinely monitored in the USA, Canada, Europe, throughout Asia, and elsewhere. Networks of in situ monitoring stations in these regions provide constant measures of air quality and help governments and concerned organizations determine how close they are to meeting their air quality goals. Precipitation chemistry monitoring also provides data on pollutants in air. Satellite remote sensing and air quality models complement the data collected by monitoring stations to form a more complete picture of air quality and can offer the advantage of consistent methods across the

globe. Stack emissions for point sources are occasionally monitored in some countries to complement the ambient concentration data collected by monitoring networks, though more often large emitters estimate their emissions using mass-based flow balances for their site and rates of activity. Numerical modeling of air pollution can be a stand-in when measured data are sparse, it can be used in conjunction with other forms of monitoring, and can help predict the effects of changes in emissions patterns. Craig et al. [13] list the outcomes of monitoring air pollution as follows:

- Ability to describe risks and detect potential future risks.
- Documentation of trends, which can demonstrate the impacts of policies.
- Inputs for the further development of models which can improve predictive capacity.
- Information necessary to prove links between pollution concentrations and effects.

Typical observations made during monitoring include concentrations in ambient air (in ppm,  $\mu\text{g}/\text{m}^3$ , particles/ $\text{m}^3$ , mol/L, etc.), rates of emission, typical concentration ranges, rates of change of concentration, and concentration profiles. It is in terms of ambient concentration that pollution standards are set. For example, the European standard for ozone is set such that the maximum concentration of ozone averaged over any 8 h period of the day should not exceed  $120 \mu\text{g}/\text{m}^3$ . Thus, it can be seen that monitoring is important to ensuring that air quality standards are adhered to, and that public health and the environment are protected.

### In Situ Monitoring

Air quality monitoring stations exist in the hundreds in some countries and in the thousands globally, covering continental areas and forming networks that contribute the bulk of information to assessments of outdoor air quality. Instruments capable of detecting specific pollutants ( $\text{NO}_x$ ,  $\text{SO}_x$ ,  $\text{PM}_{2.5}$ , etc.) draw in ambient air and accurately measure the pollutant mass per standard volume of air. This type of monitoring, referred to as in situ monitoring, results in the least uncertainty as compared to satellite remote sensing and model estimates. It is important to recognize that the in situ data



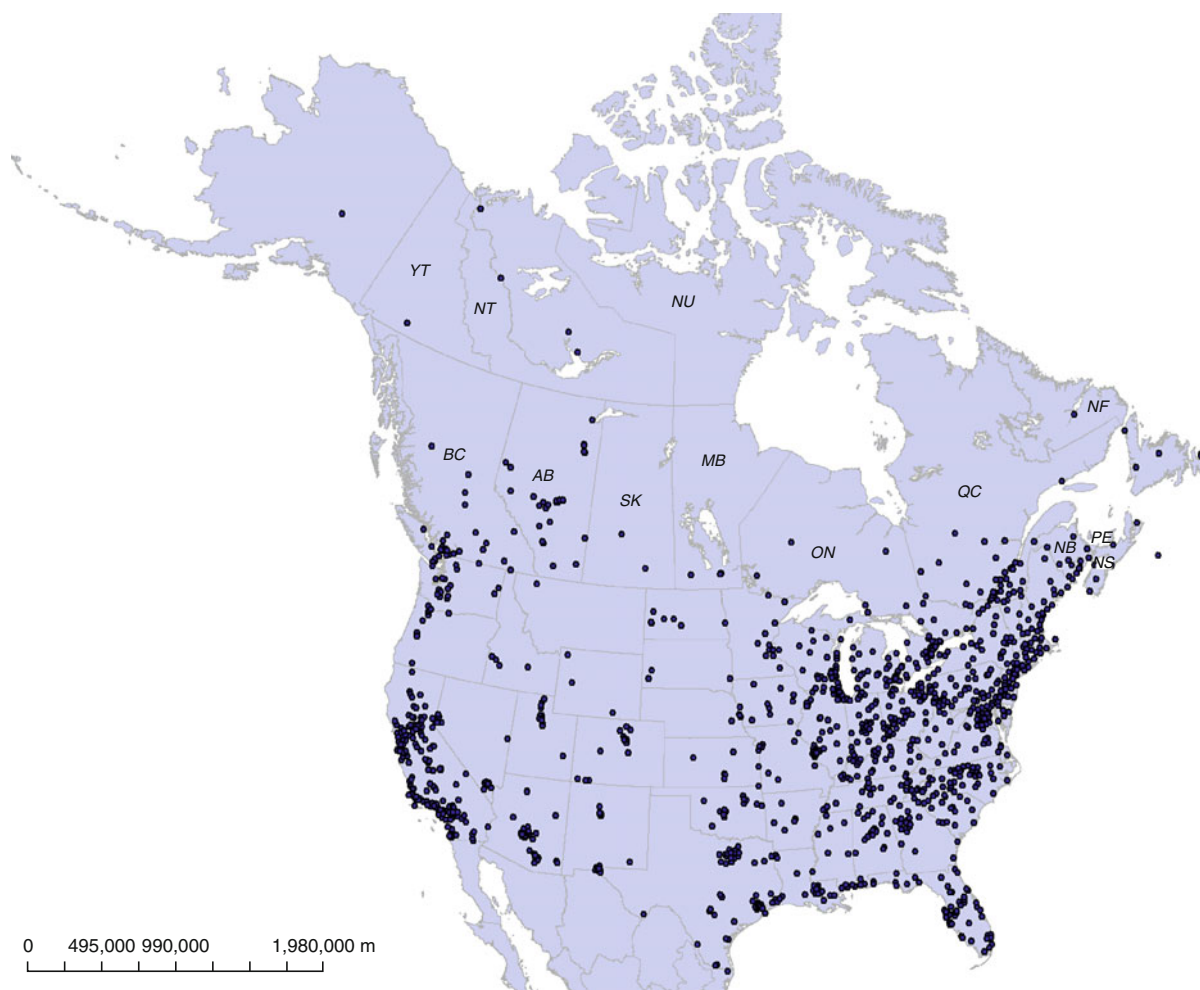
from a given monitoring location are only for the specific geographic location and, thus, during the process of site selection, the issue of spatial representativeness must be considered in the context of monitoring objectives.

Most existing stations measure near ground level ( $\sim 3\text{--}10$  m) so that air pollution at the exposure height of the general populace can be quantified. Figure 1 shows the location of ozone monitoring stations throughout North America. It is largely through such networks of stations that an area's compliance with current air quality standards is determined. Networks are also a primary tool in the tracking of progress toward air quality goals, and they provide the necessary baseline measurements for judging whether progress is

occurring [87]. Despite there being over 1,000 stations in Fig. 1, there are relatively large unmonitored areas. Network design to minimize the impact of this inevitable incomplete coverage remains an ongoing challenge and, as described below, models and remotely sensed observations are helping to fill in the gaps.

### Precipitation Chemistry Monitoring

By collecting and analyzing precipitation, stations and networks in North America (e.g., CAPMoN, NADP [76]), Asia (e.g., EANET [77]), throughout other continents, and over the Ocean [78] can collect data on air pollution. Precipitation samples can provide



**Air Pollution Monitoring and Sustainability. Figure 1**  
Ozone monitoring station locations throughout North America

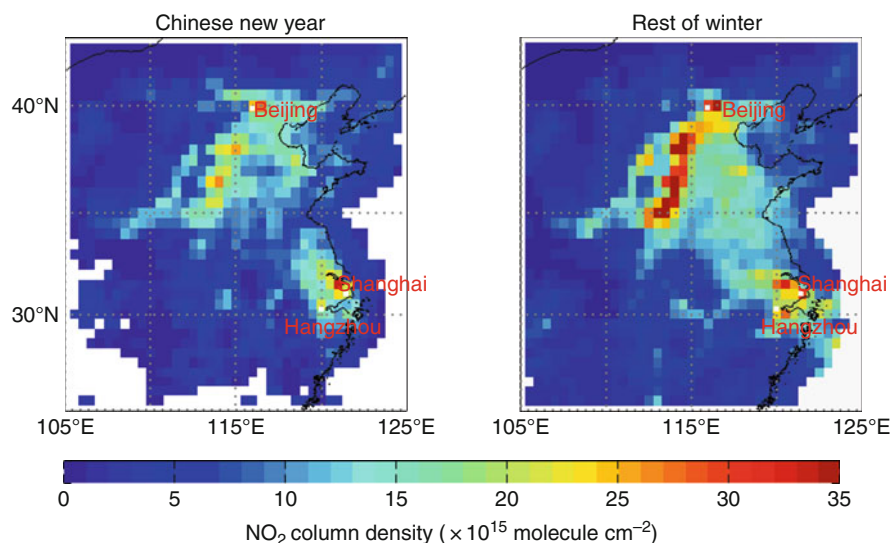
information on pH and the mass of various ions such as  $\text{SO}_4^{2-}$ ,  $\text{NO}_3^-$ ,  $\text{Ca}^{2+}$ , and  $\text{Na}^+$  [76]. These data are valuable in tracking the acidity of rain over time and in studying wet deposition. The identity and concentrations of ions discovered through precipitation chemistry monitoring can also provide information about the sources of pollutants, as in Thepanondh et al. [78]. By monitoring the concentration of wet-deposited air pollutants in precipitation, precipitation chemistry monitoring acts as a valuable element of air pollution monitoring.

### Satellite Remote Sensing

Satellites can observe electromagnetic radiation from the atmosphere and Earth's surface to infer information about concentrations of aerosols and trace gases in the atmosphere. Pollutants that can be observed by satellites include PM, tropospheric  $\text{O}_3$ ,  $\text{NO}_2$ , CO, HCHO, and  $\text{SO}_2$  [88]. Figure 2 shows satellite-observed  $\text{NO}_2$  concentrations in Eastern China.

Satellites can be used to analyze and forecast air quality, provide information on tropospheric air

pollution, and be used in conjunction with computer models. Analyses made possible by satellites include observation of events too large in scale to be appreciated using monitoring stations on Earth's surface. Such events include the difference in regional  $\text{NO}_2$  concentrations between average conditions in China and the period surrounding Chinese New Year illustrated by Fig. 2. Many spatial features related to typical emissions and those resulting from the change in emissions are better characterized and hence understood using the satellite observations. Forecasting is made possible by observation of synoptic scale meteorology. Observations of surface air quality can be combined with other observations such as those from monitoring stations and point source emissions monitoring to increase certainty of pollutant concentration measurements. Computer modeling aided by satellite remote sensing allows for pollution source apportionment by working backward from present pollutant concentrations and distributions to identify sources [88]. Point source data can be used to model forward in time to obtain estimates of future pollutant concentration. Together, satellite observation and computer modeling



**Air Pollution Monitoring and Sustainability. Figure 2**

Ozone Monitoring Instrument retrieved  $\text{NO}_2$  columns over Eastern China. All data for January 25 to March 5 were averaged for years 2005–2010. The left image shows only overpasses taken between Chinese New Year and 2 weeks after Chinese New Year. The right image shows columns from all other days during the period in question. Satellite measurement can show regional scale effects which would be difficult to discern with a surface monitoring network. Free use of tropospheric  $\text{NO}_2$  column data from the OMI sensor from [www.temis.nl](http://www.temis.nl) is acknowledged [89]

can increase the certainty of air quality estimates [85] and the attribution of pollutants to specific sources. Satellite observations can also achieve spatial resolution of pollutant concentration over an area that could only be replicated by a large number of ground-based monitoring stations, a number that could only come at great cost [85].

Another advantage of satellite observations is global uniformity. Satellites should ideally be able to measure regions equally well across the globe, in developed and developing countries, using the same measurement techniques and producing data in a uniform format. In reality, differences in surface albedo, vertical mixing, aerosol optical depth, and other factors create inconsistencies in the surface concentrations estimated from satellite data for different regions. However, once these technical challenges have been overcome, this global uniformity should facilitate simpler international cooperation in studies and provide a standard set of data for the comparison of global cities' air pollution footprints. Satellite observations also allow for the tracking of these city footprints over time, allowing for a record of the trend toward sustainability. These advantages make satellite observation an important evolving tool in global cooperation in the pursuit of sustainability. Developing methods for using satellite-derived data to provide information on sustainability should be a priority in future.

### Emissions Monitoring

Emissions monitoring provides data on the rates of emissions from pollution sources. These data make for strong inputs into air quality models [87], can ensure that emissions are not exceeding permitted values (either legally permitted or permitted with respect to sustainability), and lead to confirmation that emissions are decreasing over time in compliance with regulations. Emissions trading schemes are dependent on emissions monitoring. Figure 3 shows a Google Earth image of point sources of air pollution in North America.

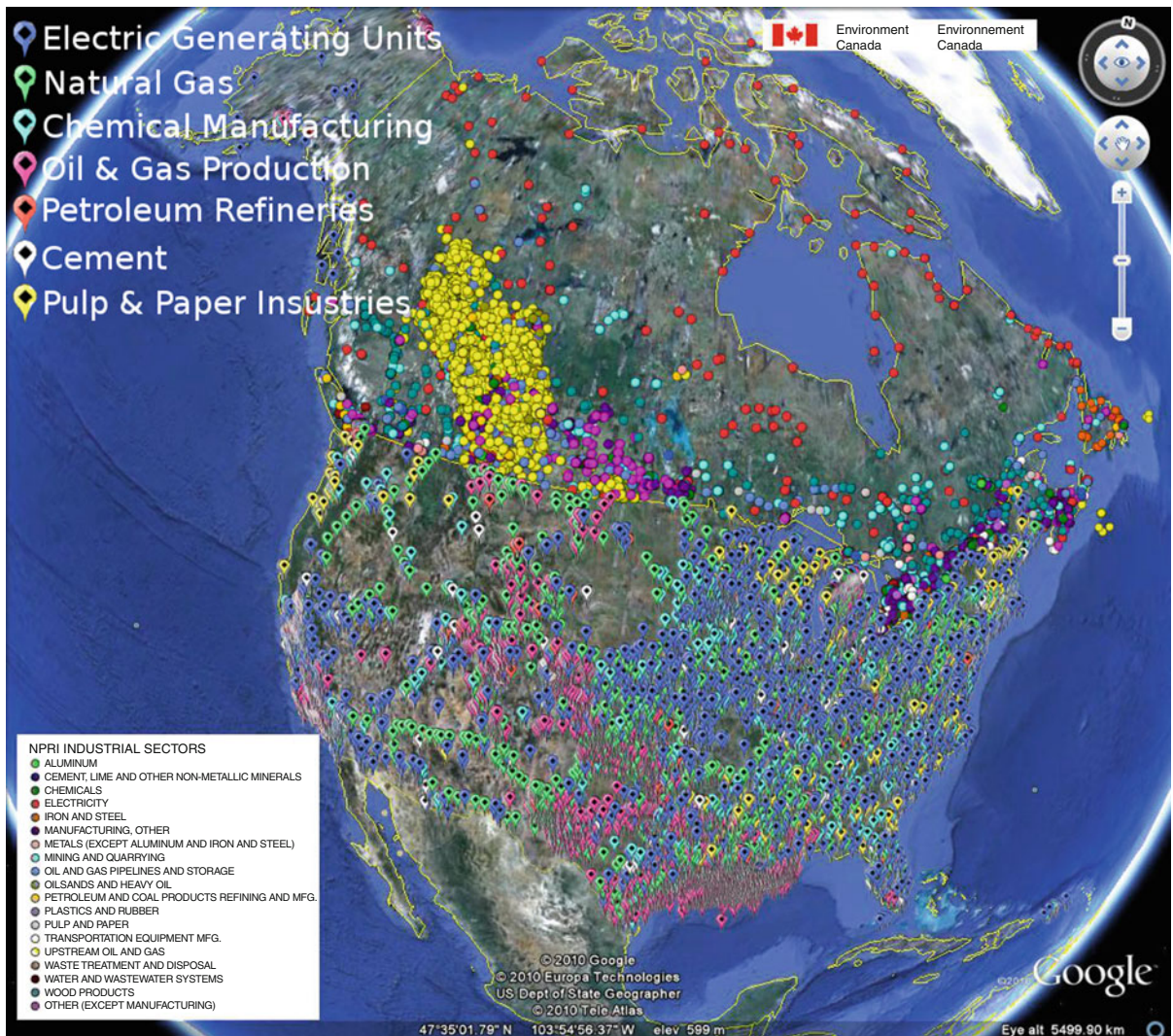
Continuous Emissions Monitoring Systems (CEMS) quantify gas or particulate emissions by taking measurements as pollution is released. CEMS are common for some major stationary sources in the USA and Europe, where their use is mandated by

legislation [79, 80]. PEMS (Predictive Emissions Monitoring Systems) use emissions factors and activity indicators (e.g., fuel use, engine types, duty cycles) in calculations that estimate emissions without measuring them. CEMS and PEMS are commonly applied to stationary sources or point sources, but emissions monitoring can also be applied to mobile sources such as locomotives [81] and ships [82]. Sources too numerous to be monitored as individual sources are often treated as "area sources" [83]. To treat sources such as residential furnaces or urban fuel stations as single sources would not be practical, so emissions from these sources are often calculated and treated as emissions not from the individual sources but from the bulk of those sources of a given type over an area. Agricultural sources such as livestock operations and emissions of  $N_2O$  from fertilizers would be considered area sources, and it is sometimes useful to think of vehicle emissions as arising from area (or line) sources. Monitoring emissions from any of these sources requires cooperation from industry and government [87]. Efforts made toward achieving sustainability will rely on accurate and timely emissions monitoring data both to inform model-based decision making and to ensure that emission reduction goals are achieved.

### Numerical Modeling

Air quality models use input from meteorological models and emissions from point and area sources to estimate air pollutant levels and deposition amounts based upon current understanding of relevant chemical and physical processes. Models can predict chemical changes in air pollutants over time [13], clarifying links between secondary pollutants and sources. Models can estimate changes in air pollution concentrations that would result from changes in practice (i.e., emissions), making modeling a valuable tool in policy design and analysis of proposed policy [87]. While originally developed mainly for this purpose, models are increasingly being used to provide optimal estimates of air pollutant concentrations over areas where data do not exist. Approaches for using models in conjunction with in situ monitoring data and satellite data are rapidly evolving [85, 86]. This practice can lead to the most complete, consistent spatial-temporal picture of air pollutant levels on the regional to global scale and is





**Air Pollution Monitoring and Sustainability. Figure 3**

Google Earth image showing point sources of air pollution in North America. USA layer available from the US EPA at <http://www.epa.gov/air/emissions/where.htm>. Canadian layer available from Environment Canada's National Pollutant Release Inventory (NPRI) at <http://www.ec.gc.ca/inrp-npri/default.asp?lang=En&n=1D892B9F-1#npri>. Both include source specific data on tonnage of criteria pollutants emitted. Legend for USA sites at *top left*, legend for Canadian sites at *bottom left*

also valuable for identifying weaknesses in current emission inventories. These “fused” data (data derived from multiple types of observations) can extend estimates of the impact of air pollution has on health and the environment, helping to better document problems and improve research. Increasing interest in providing the public with future air quality information, which

as described above can help reduce impacts, has led to models that are optimized for forecasting on a daily basis.

Regardless of their application, air quality models require evaluation to assess the reliability of their results. Any model being used to inform policy should have each of its aspects verified, including its design,

methods for representing processes, input data (e.g., emissions), and numerical methods [87]. When supported by detailed measurements, evaluation exercises also identify specific weaknesses in models and possible ways to improve future versions.

### Policy and Air Pollution Standards

Air pollution policy is a major influence on monitoring and monitoring is a major influence on policy. Policies determine legally acceptable concentrations of air pollutants. Monitoring shows which pollutants must be subject to improved standards and which areas require the most pollution-abatement planning.

Policies on air pollution centre on concentration standards and emission standards. Standards are often set based on concentrations deemed achievable using the Best Available Technology (BAT) for emissions reduction (an example of a “command and control” strategy), the effects of air pollution deemed acceptable (an example of “risk-based” strategy), or a hybrid of the two strategies [90].

While monitoring and the related science is one influence on policy and standards, it is not the only influence. Political and economic considerations also play a role and can make for situations where environmental damage is accepted despite preventability. Monitoring alone cannot deliver sustainability without policy and practices in accordance with the information it yields.

**Concentration Standards** Assessing compliance with standards is the primary goal of the majority of air pollution monitoring [91]. These standards are in the form of maximum allowable concentrations of air pollutants. Table 6 gives standards for some major world regions, which are often expressed as an average concentration over a specific time period, for some major world regions.

**Emissions Standards** Figure 4 shows how emissions reduction programs play a part in achieving concentration standards once air quality goals are established. Emission reduction programs can involve encouragement of energy efficiency, changes in practices (e.g., switching to cleaner fuels, use of emissions controls technology), and redesign of processes. They can also

include setting of emissions standards, which limit emissions from particular sources or classes of sources. For example, vehicles in Europe must not exceed the allowable emissions under the “Euro” class of emissions standards for CO, NO<sub>x</sub>, certain classes of hydrocarbons, and PM [94].

Monitoring is integral to the “Establishment of Air Quality Goals” and “Track and Evaluate Progress” stages of Fig. 4 and informs the understanding of “Current Ambient Air Quality.” Monitoring not only informs policy development, but also the assessment of policy effectiveness.

### Non-environmental Considerations in the Setting of Standards

Air pollution monitoring is one influence on policy and concentration standards, but politics, economics, and other considerations also play a role [13]. There are at times differences between what the scientific data recommend as the concentration standard and what the concentration standard is in law. The usage of “critical loads” and “target loads” in the in the Eastern Canada Acid Rain Programs is a telling example of the difference between the two terms. A critical load is what has been determined to be close to the maximum amount of pollution an ecosystem can tolerate. (In the case of Environment Canada’s Acid Rain critical loads, the aim is to protect 95% of lakes, based on surface area, from acidification beyond critical levels.). A target load, the pollutant load aimed for in policy, is the load deemed achievable once “other factors (such as ethics, scientific uncertainties, and social and economic effects) are balanced with environmental considerations.”[96]. Though the Canada-Wide Acid Rain Strategy for Post-2000 set the long-term goal of meeting the critical load, in the short term this system of standard setting determines a level of pollution slightly more than what could be sustained (the critical load) and aims to allow more pollution than that (by aiming for the target load). The critical and target load system has also been applied in areas of Europe. The different grades of Chinese air quality standards in Table 6 are another example of standards not based on the recommendations of the air pollution monitoring community or on the basis of sustainability. The different grades are assigned based on land use (e.g., residential vs industrial) and the economic realities of those areas, not on changes in the definition of



**Air Pollution Monitoring and Sustainability. Table 6** Air quality standards by world region

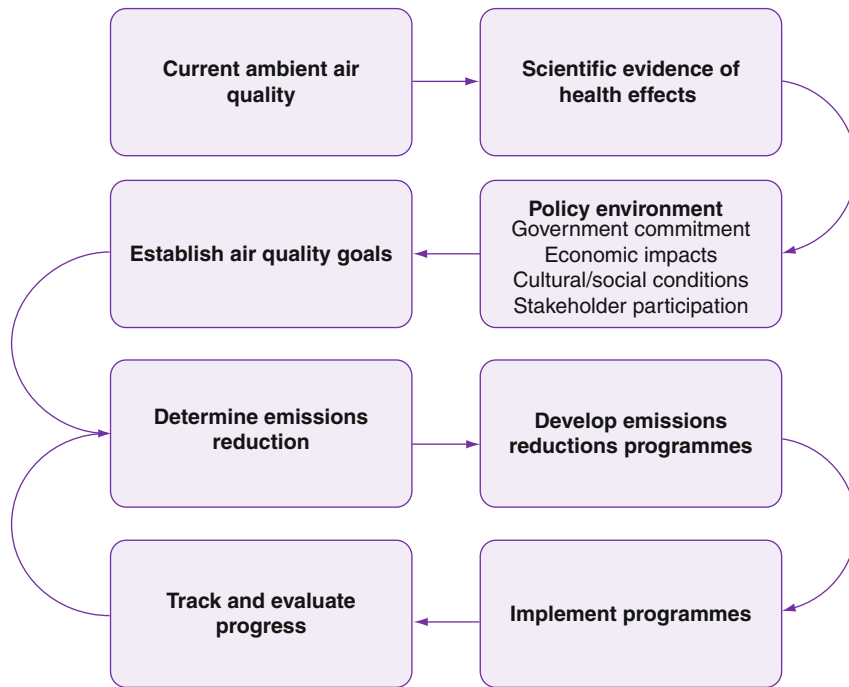
| Pollutant         | Averaging period | Standards                       |                               |                              |                              |                              |
|-------------------|------------------|---------------------------------|-------------------------------|------------------------------|------------------------------|------------------------------|
|                   |                  | Europe [92]                     | USA primary standards [93]    | China [53]                   |                              |                              |
|                   |                  |                                 |                               | Grade 1                      | Grade 2                      | Grade 3                      |
|                   |                  | Conc.                           | Conc.                         | Conc.                        | Conc.                        | Conc.                        |
| CO                | 1 h              |                                 | 35 ppm                        | 10 $\mu\text{g}/\text{m}^3$  | 10 $\mu\text{g}/\text{m}^3$  | 20 $\mu\text{g}/\text{m}^3$  |
| CO                | 8 h              | 10,000 $\mu\text{g}/\text{m}^3$ | 9 ppm                         |                              |                              |                              |
| CO                | 24 h             |                                 |                               | 4 $\mu\text{g}/\text{m}^3$   | 4 $\mu\text{g}/\text{m}^3$   | 6 $\mu\text{g}/\text{m}^3$   |
| NO <sub>2</sub>   | 1 h              | 200 $\mu\text{g}/\text{m}^3$    |                               | 120 $\mu\text{g}/\text{m}^3$ | 240 $\mu\text{g}/\text{m}^3$ | 240 $\mu\text{g}/\text{m}^3$ |
| NO <sub>2</sub>   | 24 h             |                                 |                               | 80 $\mu\text{g}/\text{m}^3$  | 120 $\mu\text{g}/\text{m}^3$ | 120 $\mu\text{g}/\text{m}^3$ |
| NO <sub>2</sub>   | 1 year           | 40 $\mu\text{g}/\text{m}^3$     | 0.053 ppm                     | 40 $\mu\text{g}/\text{m}^3$  | 80 $\mu\text{g}/\text{m}^3$  | 80 $\mu\text{g}/\text{m}^3$  |
| SO <sub>2</sub>   | 1 h              | 350 $\mu\text{g}/\text{m}^3$    | 0.075 ppm                     | 150 $\mu\text{g}/\text{m}^3$ | 500 $\mu\text{g}/\text{m}^3$ | 700 $\mu\text{g}/\text{m}^3$ |
| SO <sub>2</sub>   | 24 h             | 125 $\mu\text{g}/\text{m}^3$    | 0.14 ppm                      | 50 $\mu\text{g}/\text{m}^3$  | 150 $\mu\text{g}/\text{m}^3$ | 250 $\mu\text{g}/\text{m}^3$ |
| SO <sub>2</sub>   | 1 year           |                                 | 0.03 ppm                      | 20 $\mu\text{g}/\text{m}^3$  | 60 $\mu\text{g}/\text{m}^3$  | 100 $\mu\text{g}/\text{m}^3$ |
| Lead (Pb)         | 3 months         |                                 | 0.15 $\mu\text{g}/\text{m}^3$ |                              |                              |                              |
| Lead (Pb)         | 1 year           | 0.5 $\mu\text{g}/\text{m}^3$    |                               |                              |                              |                              |
| O <sub>3</sub>    | 1 h              |                                 |                               | 160 $\mu\text{g}/\text{m}^3$ | 200 $\mu\text{g}/\text{m}^3$ | 200 $\mu\text{g}/\text{m}^3$ |
| O <sub>3</sub>    | 8 h              | 120 $\mu\text{g}/\text{m}^3$    | 0.075 ppm                     |                              |                              |                              |
| Benzene           | 1 year           | 5 $\mu\text{g}/\text{m}^3$      |                               |                              |                              |                              |
| PAHs              | 1 year           | 1 ng/m <sup>3</sup>             |                               |                              |                              |                              |
| PM <sub>10</sub>  | 24 h             |                                 | 150 $\mu\text{g}/\text{m}^3$  | 50 $\mu\text{g}/\text{m}^3$  | 150 $\mu\text{g}/\text{m}^3$ | 250 $\mu\text{g}/\text{m}^3$ |
| PM <sub>10</sub>  | 1 year           |                                 |                               | 40 $\mu\text{g}/\text{m}^3$  | 100 $\mu\text{g}/\text{m}^3$ | 150 $\mu\text{g}/\text{m}^3$ |
| PM <sub>2.5</sub> | 24 h             |                                 | 35 $\mu\text{g}/\text{m}^3$   |                              |                              |                              |
| PM <sub>2.5</sub> | 1 year           | 25 $\mu\text{g}/\text{m}^3$     | 15 $\mu\text{g}/\text{m}^3$   |                              |                              |                              |
| Arsenic (As)      | 1 year           | 6 ng/m <sup>3</sup>             |                               |                              |                              |                              |
| Cadmium (Cd)      | 1 year           | 5 ng/m <sup>3</sup>             |                               |                              |                              |                              |
| Nickel (Ni)       | 1 year           | 20 ng/m <sup>3</sup>            |                               |                              |                              |                              |

sustainability from area to area. Policy-makers consider much more than what is sustainable when setting standards, and when doing so, the tendency is to make standards less stringent, meaning sustainability is sacrificed further.

Polluters can also play a role in affecting air quality standards. There are concerns that setting stringent air pollution standards will cause polluting industries to relocate to “pollution haven” countries where standards are more lax, though the idea that moving to one of these havens actually benefits businesses is dubious [97]. Fear of such moves can erode the economic

base of a policy-maker’s purview, providing an incentive to weaken standards. Fredriksson and Gaston [98] review a number of cases in which environmental policy was weakened by business lobbies representing polluters. Clearly, pollution policy making is the result of many disparate influences. Amid social, political, and economic debates involving varying interests, the goal of sustainability is often minimized or even lost.

**Violations of Concentration Standards** Unfortunately, standards are targets, not guarantees of



**Air Pollution Monitoring and Sustainability. Figure 4**

Steps in the emissions reduction cycle including policy lead-up. Figure is based on [95] and [13]

maximum concentrations. As of 2004, 20.2% of Chinese cities exceeded Grade III standards, though this rate was an improvement over the 53.8% of cities exceeding Grade III standards in 1991 and was part of an overall trend of decreasing concentrations [53]. In the case of the USA, it must be noted that the trend in average concentrations of most criteria air pollutants is downward [99]. However, in the USA, in 2007, 21% of sites monitoring  $\text{PM}_{2.5}$  exceeded the 24 h standard of  $35 \mu\text{g}/\text{m}^3$ , 23% of sites monitoring lead showed concentrations in excess of the  $0.15 \mu\text{g}/\text{m}^3$  standard, and 57% of sites monitoring  $\text{O}_3$  exceeded the 0.075 ppm standard [95]. The UK's ozone standard is exceeded several times each year [83]. In recording exceedences of standards, air pollution monitoring demonstrates the need for continued improvements in emissions abatement and shows where the need is greatest. It also demonstrates that standards and monitoring alone cannot lead to sustainability. Strong enforcement of standards must make air pollution a consideration in the development and uptake of technologies.

**The Effects of Air Pollution Science on Policy** Air pollution policy, like all environmental policy, is dependent on research and the state of science. If science has not developed the discourse to describe an air pollution issue, there are no grounds or even terms to support policy-based mitigation. For example, it is believed that different species of air pollutants may act together to produce more severe health impacts than either species would produce if acting alone, but the current understanding of synergistic interactions is inadequate to support this level of policy development [101].

Air pollution science is also critical in identifying the most significant contributors to pollution [101] and can use this and other information to recommend optimally cost-effective solutions to air pollution problems [102]. The concept of critical load is an example of a scientific tool that aids air pollution problem solving [102]. It is also an example, as discussed above, of scientific knowledge being used as a framework but not being fully respected by the political process. While science is the basis of policy's discourse, as well

as the basis for policy evolution, there is room for improvement in how science and policy cooperate in achieving sustainability.

### **Policy and Air Pollution Standards: Conclusion**

Policy sets standards by which the results of air pollution monitoring are judged. Concentration standards indicate accepted levels of air pollution in ambient air, and emissions standards are measures of how much any source can contribute to air pollution. Political, social, and economic considerations often influence standards and result in more lax standards than science or sustainability would recommend. Even in these cases of standards not consistent with sustainability, standards are often exceeded. Much of air pollution monitoring is designed to assess compliance with standards [91], but neither standards nor standards-oriented monitoring alone will ensure sustainability.

The current air pollution monitoring establishment can contribute to ultimate sustainability through ongoing scientific assessment in support of increasingly appropriate concentration standards. Monitoring and the related research can show, and have shown, that in many cases lower ambient air pollution concentrations are possible, necessary, and cost effective. Thus, new policy and better standards are encouraged through the results of monitoring. Translating air pollution monitoring into policies that support sustainability may be a considerable challenge. Hurdles may include political and economic dampening of the recommendations of science, and ongoing violations of existing standards. However, progress is essential as without monitoring sustainability becomes virtually impossible to verify. Further, if monitoring is to verify sustainability, it needs to be based on goals and policies that balance the social, environmental and economic aspects of sustainability. The monitoring community can play an important role in establishing such holistic standards for air quality.

### **Past and Ongoing Success in Air Pollution Mitigation**

In general, OECD countries have seen concentrations of most major (or criteria) air pollutants decrease since the introduction of pollution management programs. One of the biggest victories has been sulfur

emissions, which were the focus of acid rain reduction programs. Sulfur emissions, again, in OECD countries, are much lower than they were in the past [6]. Acid rain is decreasing in the USA [95], Canada [99], and elsewhere. NO<sub>2</sub>, thanks largely to the use of catalytic converters in vehicles, is on the decline in most European cities [6], throughout the USA [95], and Canada [100]. Limitations on lead in gasoline have led to substantial reductions of lead in air. Concentrations of lead are consistently below the WHO concentration guidelines throughout Western Europe and, in Canada, lead is almost undetectable in the atmosphere [6]. Regulations controlling particulate were the most cost-effective regulations the US government put into force [43], and human life expectancy has increased as air quality has improved in the USA [44].

Countries managing air pollution have seen a decline in ambient HAP concentration, with benzene, 1,3-butadiene, styrene, xylenes, and toluene concentrations decreasing by at least 5% per year in the USA from 2000 to 2005 [95]. A major factor in this success was the control of emissions from chemical plants, dry cleaners, coke ovens, and incinerators. The continued reduction in CFCs discussed above is a landmark success in air quality policy and international policy in general.

In many cases, serious efforts to reduce the presence of a pollutant in the environment can succeed. Air pollution monitoring brings problems to attention so that these efforts can be made. It also tracks the impacts of pollution mitigation strategies, indicating success when it is achieved and allowing for comparison of the effectiveness of different strategies. If not for monitoring, the successes mentioned here would be impossible to quantify or prove. Adaptive management strategies would also be impossible.

### **Outstanding Problems and Challenges**

The successes mentioned above are not unmitigated. pH and ANC in ecosystems damaged by acid rain are not recovering in proportion to decreases in acidifying emissions [46]. The Ozone layer's recovery is not proceeding in proportion to decreases in CFC emissions but is being retarded by other anthropogenic air pollutants [72]. Catalytic converters used to decrease NO<sub>2</sub> emissions cause a net increase in fuel use, prompting Fenger [6] to comment we may have mitigated one

problem by aggravating another. Reduction in NO<sub>2</sub> emissions have not led to the reductions in tropospheric ozone that were hoped for [11].

The story of lead as a gasoline additive is an example of how reduction of a pollutant does not always result in greater sustainability. With lead no longer an acceptable fuel additive, methyl *tert*-butyl ether (MTBE) was used to perform its function. However, MTBE was later found to be a powerful contaminant of water and a likely carcinogen [103]. Now bio-ethanol is being discussed as a less poisonous fuel octane booster than MTBE, but it introduces its own risks of soil and groundwater contamination and can lead to increased emissions of NO<sub>x</sub> and organics [104], not to mention stresses on global food supplies and agricultural resources (i.e., land and water). While legislation such as the USA's Clean Air Act was successful in reducing emissions from gasoline, it did not eliminate the environmental threat of fuel oxygenates. More holistic systems and policies, rather than the reactionary phasing out of particular pollutants, are necessary to truly solve the problem. Further, when implementing such policies, we need to anticipate potential associated consequences and be prepared for follow-up monitoring so that we can detect any resulting impacts.

In addition to the problems remaining after emissions and ambient concentrations of pollutants have been reduced, there are instances where reductions in concentrations elude the efforts of policy, technology, and behavioral changes. Over half of the population of the USA lives in counties where air quality standards are exceeded [95]. O<sub>3</sub> in populated areas such as those in Southern Ontario is continuing a long-term trend of increasing concentrations [105] despite decreases in concentration of precursors, leading to questioning of the fundamental scientific assumptions surrounding tropospheric ozone concentrations [106]. Background ozone is increasing in parts of the northern hemisphere, and this increase may be linked to intercontinental transport of pollutants [107]. Ozone and PM concentrations also continue to be challenges throughout the USA [95].

Despite the effort and advances made in research, policy, and environmental practice, air pollution is still a health problem, a visibility problem, and an economic problem. Further, only weak links exist between

air quality objectives and sustainability. Concentration and emissions standards are not set to serve sustainability. Air quality objectives that will allow the eventual recognition of sustainability once it has been achieved, or even how far away sustainability currently is, are lacking. The biosphere is still being damaged by air pollution. Acid rain is declining in developed countries, but is still beyond the critical load, is increasing in some regions, and totally unmonitored in others. Thus, despite progress in recent decades, there is little reason to believe that we currently have a sustainable burden of pollutants in our atmosphere. Thus, there are considerable challenges remaining for the air pollution monitoring community in relation to sustainability.

### Current Air Pollution Monitoring and Sustainability

Diverse techniques contribute to the monitoring of pollution in the atmosphere. This monitoring allows for air quality advisories, research into the nature and sources of air pollution, evaluation of the conformity with standards, and assessment of the need for additional or alternate regulations or policies. All of these outcomes of air pollution monitoring reduce the burden of air pollution on human society and the environment. Unfortunately, what is learned and understood is not proportionate to action taken. Pollution is still eroding the capacity for sustainability. Improvements on all fronts are necessary, including improvements in air pollution monitoring, the associated science, practices, the thinking that underpins them, and the social response they receive. Determination of what constitutes a sustainability standard or threshold will also require further work.

### Future Directions: Monitoring to Promote Greater Sustainability

In redesigning the air pollution monitoring community and the larger air pollution establishment for sustainability, there are a number of changes to be made to current practices. There are both broad, or institutional, changes and specific changes to be made. The institutional changes will be more difficult to affect in that they require overcoming current inertia. These changes are discussed immediately below. Specific changes are discussed in the [“Monitoring More](#)

[Effectively](#)” section. Both types of change will help the science of air pollution monitoring encourage sustainability.

### Changes in Institutional Practices

Institutional changes that would make air pollution monitoring more relevant to sustainability include the expansion of the conventional goals of air pollution monitoring, working across intellectual disciplines, and involving the general public in the monitoring process.

**Expansion of Goals** Air pollution monitoring practices are largely determined by the goals underlying air pollution mitigation strategies. When assessment of health impacts is the goal, more monitoring happens in populated urban areas. The hierarchy of goals in current air pollution mitigation tends to prioritize human health, then visibility, then ecosystems [54], then possibly the prevention of economic damage from air pollution, or even climate stability in rare cases. Often studies are oriented toward only one of these goals.

Sustainability requires holistic thinking and if sustainability is the aim, prioritizing based on a hierarchy of problems alone can reduce the potential benefits. Prioritizing air pollution impacts immediately affecting humans (e.g., health effects) over environmental impacts is self-defeating. In the long term, it is possible that the loss of ecosystem services arising from these lower-priority environmental effects may be the greater consequence to humans [54]. Holistic thinking would take into account many more goals in air pollution mitigation than health, visibility, ecosystem damage, economic loss, and climate. A holistic system would involve monitoring air pollution with goals such as the development of natural capital, resilience in human and natural systems, environmental justice, improved urban form, and accelerated learning and research. Studies with these goals in mind have been performed infrequently, and they are fertile ground for new research.

*Natural Capital* Natural capital is a metaphor used to describe the natural features of ecosystems that do work or provide services for humans and the biosphere. It is useful in that it draws attention to

these features and their indispensable services such as clean air, potable water, food, and building materials while drawing comprehensible analogs to economic concepts. Pollution that makes air a health risk, destruction of the ozone layer by air pollutants, acidification of ecosystems via acid rain, and alteration to the climate by air pollutants such as black carbon aerosol are all examples of air pollution destroying natural capital. All of Guderian’s categories [54] (“[Damage to the Biosphere](#)” section) are ways in which air pollution depletes natural capital. Natural capital is often extended to include the capacity of systems to assimilate waste, such as the atmosphere’s capacity to assimilate air pollution, and the life support functions provided by ecosystem components, such as the ozone layer’s UV blocking capabilities [108].

The assessment of the cost-effectiveness of the US Clean Air Act, as mentioned, gave a mid-value of the benefits of the Act of \$22.2 trillion but could not include the further benefits of preservation of natural capital [19]. All production and consumption in the human economy depends on natural capital [108, 109]. Minimum estimates of the value of natural capital and ecosystem services exceed the value of the human economy [110, 111]; yet, the discourse for the quantification of natural capital has yet to be fully developed. Furthermore, efforts to monitor and quantify pollution’s impact on natural capital are rare [112], making the inclusion of natural capital loss in the Clean Air Act’s benefits or in the cost-benefit analysis of new air quality policy options near impossible. Data on natural capital are limited and thus far unsatisfactory, and types of natural capital, such as types more vulnerable to air pollution versus types more vulnerable to other damages, have yet to be disaggregated [113]. Assessment of loss of natural capital must become a priority for the air pollution monitoring community, as should the categorization of types of natural capital vulnerable to air pollutants. Mechanisms to ensure that natural capital is fairly recognized in policy development need to be installed at the international level.

Critical natural capital, natural capital that cannot be lost without irreparable terminal damage to ecosystems and thereby similar damage to the human economy, must also be considered. This critical natural capital cannot be lost or exchanged and, therefore, is literally priceless [108]. Criticality of natural capital can



be judged in terms of source and sink functions (sources of ecosystem services and pollution sinks) and life support functions [114]. Examples of critical natural capital would be keystone species or keystone processes [108], which can be affected by air pollution. There is a certain level of natural capital that must be maintained to preserve the biosphere and hence humanity. Unfortunately, where this level lies is not yet understood, especially in terms of atmospheric aerosol loading and chemical pollution [115], two areas directly relevant to the air pollution monitoring community. Air pollution monitoring and air pollution science must be oriented toward defining where atmospheric natural capital becomes critical, how air pollution is damaging natural capital, and the quantification of natural capital savings from emissions reductions. Air pollution monitoring could also provide data on the development and redevelopment of natural capital.

*Resilience* A sustainable system must be resilient to external shocks and major changes, but the socio-ecological concept of resilience encompasses more than an expectation of a return to business as usual. Socio-ecological resilience is not only the ability of a system to retain basic functions despite a shock. It is also the capacity for renewal, reorganization, and development [109]. Air pollution monitoring should accept the possibility of major system changes and monitor for gradual atmospheric change as early warning signs of upcoming system shifts [114, 116]. Early warnings would allow for preparation for adaptation at worst, possibly allow for avoidance of the worst consequences of system shocks, and expose opportunities to take advantage of upcoming changes at best. The development of appropriate atmospheric indicators of ecological system shifts would be a major research project and require much innovative air pollution monitoring. One aspect of such indicators would be placing the concentration of air pollutants in the context of the capacity of systems to continue to absorb more pollution without major changes in their behavior [114]. Indicators could also include how the dynamic capacity (ability to reorganize around changes) of environmental sinks (e.g., the atmosphere) are affected by continued pollution [114]. These types of indicators would be along the

lines of damage-oriented metrics used in life-cycle assessment methods.

Resilient systems are contrasted with vulnerable systems, in which small changes can cause disastrous consequences. The resilience of the human population is a sustainability-relevant metric to be developed. Resilience in this population comes under fire when exposure to pollution increases human vulnerability to disease, as has been shown to be a likely outcome of exposure to some air pollutants [117, 118]. Monitoring to expose the development of vulnerabilities would be a useful endeavor in the promotion of resilience. It would also be useful to show how the allocation of the most polluted land to the most vulnerable populations (babies, young people), as in Mitchell and Dorling [119], and the associated increases in asthma, other respiratory illnesses, and allergies are affecting the ability of humans to reorganize around potential atmospheric changes. Development of such indicators will rely on air pollution monitoring data.

*Socio-ecological Justice* The social justice of environmental protection is a fundamental tenet of sustainable development [119]. Environmental justice is concerned with equal access to a clean environment and equal protection from risks associated with environmental impurities [119]. Currently, economically disadvantaged communities are often more exposed to air pollution than more affluent communities [117]. Poorer individuals are more likely to live near roadways, resulting in greater exposure [117]. In their study of London, MacKerron and Mourato [120] observed that communities with the least car ownership are exposed to the highest concentration of pollutants, while the cleanest air tends to be in communities where car ownership is greatest. Marshall et al. [121] shows that lower-income neighborhoods tend to have higher walkability (resulting in less auto emissions), but also higher NO concentrations, while higher income neighborhoods tend to have lower NO and O<sub>3</sub> concentrations. In these cases, the injustice is that the people polluting the most impose the cost of their pollution on those polluting least. Identification of such injustices should be a focus of air pollution monitoring.

Routine increased exposure itself may enhance the impacts of air pollution on socioeconomically

disadvantaged groups by weakening normal biological resistance [117] or promoting atherosclerosis [118]. Housing stock in poorer neighborhoods can have higher rates of off-gassing of some allergens and expose occupants to other risk factors of asthma sensitization and exacerbation [117]. These phenomena require greater documentation by air pollution monitoring agencies so that they can be addressed.

Often young families cannot afford to live in more affluent neighborhoods and must purchase lower cost housing. This situation has led to findings such as those of Mitchell and Dorling [119], that pollution in London is most concentrated where young children and their parents live. London wards with a high proportion of all babies in the city tend to have much higher than average concentrations of NO<sub>2</sub>. This injustice is exposing pregnant women, children, and babies (segments of the population very vulnerable to the health effects of air pollution) to greater air pollution concentrations than those more resistant.

Much environmental justice work has been done in the USA to attempt to redress the effects of the frequent location of landfills and polluting industries in black and indigenous communities. A lack of empirical evidence led to failure of many of the class actions meant to address environmental injustice [119]. Air pollution monitoring should be used to supply empirical evidence where environmental injustice exists. As O'Neill et al. [117] state, "Science and public policy would benefit from additional research that integrates the theory and practice from both air pollution and social epidemiologies to gain a better understanding of this issue."

*Improved Urban Form* Design of urban form can mitigate exposure and reduce emissions [6]. Urban sprawl results in increased automobile dependency and the associated emissions [122]. Many studies have witnessed the negative correlation between urban density and per capita transportation emissions [123]. Frank et al. [124] makes the connection between neighborhood walkability and reduced vehicle emissions. The US EPA has acknowledged the relationship between urban form and emissions through its program of offering urban nonattainment areas emissions credits for "smart growth"

development practices [123]. Conventionally, air pollution mitigation has been technology oriented, and Stone [123] claims that the EPA's program of emissions credits for intelligent urban design is an acknowledgement that the technological approach alone is not sufficient to solve air pollution problems. This recognition reinforces the point that sustainable solutions must be holistic and that no single avenue leads to sustainability. Air pollution problems cannot be solved by dealing with the emission sources alone, there must be systemic redesign of the landscapes that create demand for emitting activities.

The air pollution monitoring community can play a role in urban design for emissions reduction. One major contribution would be taking stock of various types of urban forms that exist and the contribution of those forms to emissions. Tang and Wang [125] look at the four different urban forms on the Macau peninsula and associate form with emissions and air pollution. Observations connecting road width, road curvature, and congestion with emissions were made, and it was seen that features such as urban canyons can lead to increased pollution in areas of cities least responsible for total emissions. Other features Tang and Wang [125] recommend for study include street orientation in relation to wind direction, interactions between urban form and wind speed, building design, green spaces, and traffic control schemes. All of these aspects of urban design require further research to determine their impact on air pollution, while data from monitoring will help to substantiate the understanding thereby derived.

Designing urban areas for walkability and public transit will also impact emissions of air pollution. PODs (Pedestrian-Oriented Development) and TODs (Transit-Oriented Development) such as those discussed in Newman et al. [126], can lower transportation energy demand and thereby reduce emissions from vehicles. A study of walkability in Vancouver showed that urban design had a substantial impact on walkability and air pollution exposure, and proved an effective method for connecting air pollution monitoring to urban design and possible future development [127]. Designing for less dependence on automobiles is a very effective strategy for using urban form to mitigate air pollution, and such design has other sustainability benefits such as improved health from physical

activity [127]. Of course, designing for urban density will result in more instances of exposure to urban air pollution, providing additional incentive to minimize emissions.

The air pollution monitoring community can also contribute to improved urban form by uniting air pollution models with urban design models. Involving the air pollution monitoring community in an integrated design process for cities could lead to reductions in emissions and exposure in existing areas by designing new developments to decrease the negative impacts of stagnant air within urban canyons. A stronger air quality perspective on urban design could result in more sustainable places to live.

*Acceleration of Learning and Research* The majority of investment in air pollution monitoring is directed not toward research or improvement of research capabilities but toward assessing compliance with existing air pollution concentrations standards [91]. The focus on compliance to the exclusion of other objectives reduces the value of data acquired in that the data could also be used for source apportionment, chemical transport modeling, assessment of background levels, and relationships between pollutants and other effects [91]. The link between air pollution and urban form could also be better established if available data were used for that purpose. If priority continues to be given to compliance over other uses of data, few of the other goals discussed herein will be easily pursued. Chow and Watson [91] give specific practices that can be implemented to shift emphasis toward research, including calibrating instruments to detect lower concentrations, decreasing time resolution of sampling, and reporting data to the level of detection rather than rounding to convenient orders of magnitude. Policy recommendations are also made, including periodic reviews of instrument performance. Acceleration of learning, of which research is a fundamental part, should be more of a focus in routine monitoring. Thus resources need to be allocated to ongoing analysis and interpretation of monitoring data.

Folke et al. [116] discuss the relationship between adaptive capacity and sustainability. To lay claim to sustainability, a society must be able to manage adaptively to socio-ecological changes. This capacity, Folke

et al. [116] argue, requires mechanisms that allow governments to learn, meaning that the institutional capacity to learn must be optimized. It is not optimized in the current situation as described by Chow and Watson [91], wherein data are primarily oriented toward fulfilling routine functions. Adaptive capacity also requires spaces for institutional collaboration [116]. Data that are conventionally less informative than they could be, as air pollution monitoring data commonly are [91], are not optimally suited for collaborative purposes. The case of the need for fine spatial resolution in the MacKerron and Mourato [120] study of socioeconomic distribution of air pollution is an example where routine data would not have served the study's interdisciplinary purpose. The capacity to adapt requires not only strong data, but the ability to extract maximum information from that data. Adaptive capacity, and therefore sustainability, would be served by the improvements suggested by Chow and Watson [91], and like improvements concerning the acceleration of learning.

Chow and Watson [91] and Folke et al. [116] are effectively calling for continual improvement in capacity to learn on the job. In doing so, they are highlighting two essential components of sustainability: learning and continual improvement. Routine monitoring must cease to focus solely on compliance. It must ensure that research and improvement in the capacity for research are served. Clearly, development of best practices to ensure sustainability needs to be a priority in this learning.

*Expansion of Goals: Conclusion* The conventional goals of air pollution monitoring influence monitoring practice. Practices thus serve these goals more than others. There is a need to expand the set of goals and expand practices accordingly. Monitoring to understand impacts on natural capital is not only essential to sustainability but will improve the economic future of humanity. Resilience is a concept the air pollution monitoring community can study and apply further. The environmental injustice of air pollution's distribution must be addressed for the sake of justice and for the health of younger populations. Given the strong links between air quality and urban form, air pollution issues should not be ignored in the design of cities nor urban form

ignored in pollution studies. Acceleration of learning and research have the potential to increase the efficiency of monitoring as well as improve understanding of air pollution's effects and optimal concentrations. All of these goals must be adopted if air pollution monitoring is to holistically pursue sustainability, and sustainability will continue to be elusive if approaches are not holistic.

**Multidisciplinary Work** Attitudes fostering sustainable levels of air pollution in the coming decades will value contributions from intellectual disciplines other than air pollution science. Urban planning work can mitigate exposure to air pollution, and it is clear that improved energy efficiency will reduce emissions from fossil fuels [53, 73]. Folke et al. [116] call for consultations with communities with local ecological knowledge [116] such as neighborhoods and Aboriginal networks. Yi et al. [53] comment that sustainable consumption and production will reduce air pollution, which suggests collaboration with business. Agricultural practices are major sources of air pollution [73], suggesting collaboration opportunities with the agricultural sector. Further, collaboration between air pollution and climate change mitigation efforts can deal with major air pollutants that are also contributors to imbalanced radiative forcing [13, 73] while discouraging the implementation of climate strategies that increase air pollution [73]. Failing to collaborate will make air pollution research more difficult and less rewarding, while collaborating will open new areas that will inform monitoring priorities.

**Public Involvement** Public involvement and consultation with the general public can be of substantial benefit to science but are rarely taken advantage of. At some levels, there is an aversion to such involvement as illustrated by a survey taken by the British Royal Society which showed that 20% of scientists believed that those scientists seen to engage with the public were less well regarded by their peers [128]. Much value is overlooked when air pollution scientists do not engage with communities they intend to monitor, including sophisticated understanding of local sources of air pollution, how local air pollution changes with weather, and health effects being experienced (even if the sophistication of the knowledge is not of the same

kind that air pollution scientists generally regard as sophistication) [120]. Citizens often have or believe they have information that monitoring bodies do not have and, even if this information is incorrect, it may affect the ability to implement policy based on results from air pollution monitoring [129]. Thus, it is a good idea to at least understand how a community feels about its local air pollution before policy, and possibly even monitoring campaigns are designed. When communities do take air quality problems seriously, opportunities for progress on air quality issues arise [13].

Yearly [129] performed two case studies on air pollution and public involvement that illustrate the value of community input. The first was in Sheffield, UK, where the local government had implemented a computer model to study the local air quality. Public consultation showed that the public of Sheffield was not satisfied with the model [129]. The public criticized the model on several points, and councilors and researchers felt that the community's and citizen's groups' comments were much more detailed, reasoned, and well supported than expected [129]. Comments were acknowledged by those running the model as insightful and well founded [129]. Issues added to the agenda by the suggestion of the citizen's groups included an inquiry into the extent to which the local council was affected by what the model presented, and consideration of the "value-for-money" associated with the modeling strategy [129], as opposed to a pollution mitigation strategy.

The second case study was in Bristol, UK. Community members were asked to use markers to develop a color-coded map of Bristol that would reflect typical air pollution in the city. Not only was the resulting map a reasonable reflection of actual air pollution in Bristol, there was valuable information to be found in a comparison of Bristol's air pollution model and the community's map in areas where the two disagreed [129]. Both studies were valuable and acted as a kind of extended peer review on the policy-related science [129].

Development of scientific material for public consumption forces scientists to develop understandings of other fields, and does not remove rigor from the work but does demand skill [130]. Communication with the public is a sustainability issue in that emissions reductions recommended by the results of monitoring are easier to implement with an informed public

supporting policies they feel they helped create [130]. Alternatively, an ill-informed public can pressure governments to adopt policy for the wrong reasons, potentially leading to unsustainable environmental impacts. O'Neill et al. [117] recommends that research not only be communicated to the public, but that air pollution research should be developed in collaboration with the affected communities.

Only rarely when scientists are asked about interaction with the public do they imagine anything other than a public lecture [130]. This mistake is an extreme example of the idea that one way communication is the only way to involve the public. Sather et al. [131] shows that the public can be directly involved in research activities. The Sather et al. [131] study covered a large area (~25,000 km<sup>2</sup> of the Dallas-Fort Worth Metroplex) and help from volunteers, including volunteer EPA personnel, 4-H club members, Master Gardeners, and retired farmers, was required. At *each* network site, volunteers were able to correctly operate instruments and collect data of quality. Such involvement is surely replicable in other settings, would increase the range and cost-effectiveness of monitoring, and would have positive impacts on a community's will to reduce air pollution. In addition to sharing of information, public involvement can result in substantial, competent aid in air pollution monitoring. Widespread, hands-on public involvement would be one aspect of a more effective effort to communicate risks associated with air pollution to the public, as called for in Craig et al. [13].

**Institutional Changes: Conclusion** In order to accelerate the progress of the air pollution monitoring community toward sustainability, there are changes the monitoring community must make. Goals must be expanded beyond the conventional goals of human health, visibility, and documentation of ecosystem damage. Work must cross the boundaries of conventional intellectual disciplines to further develop the implications of monitoring data and the tools and metrics used in monitoring. Public involvement will improve research goals in many cases, bring public support to air pollution mitigation and, in some cases, aid the research process itself. Making changes in these directions will help to promote sustainability in air pollution monitoring.

## Monitoring More Effectively

Changes that can be made to standard monitoring practice discussed here are based on current trends or suggested modifications to practices. They include development of the understanding of multipollutant impacts, monitoring for universal consolidated sustainability metrics, improvements to the use of satellite observations (partially through new instrumentation), use of design methodologies for monitoring networks, and removing some of the oversimplifications routinely made in conventional monitoring. These changes can form initial steps toward improving the air pollution monitoring community's contribution to sustainability.

## Multipollutant Monitoring and Science

Health risks from air pollution are almost exclusively assessed on a single pollutant basis [101]. Health risks from, for example, ozone, are assessed separately from health risks from NO<sub>x</sub>. In reality, however, people are exposed to many pollutants at one time. Pollutants can have synergistic effects that exacerbate health impacts beyond the sum of the health impacts they would have individually [101]. Air pollution research and the US EPA are moving toward multipollutant research and legislation, but available data are inadequate to create legislative models [101]. To develop the required data, more and better monitoring, as well as improved exposure modeling and data processing are required [101]. The necessary changes will take time; however, development in this direction is required for accurate assessment of the health impacts of air pollution.

## Design for Consolidated Sustainability Metrics

Environment Canada's Air Quality Health Index (AQHI) and the US EPA's Air Quality Index (AQI) are examples of consolidated metrics. They incorporate a number of data on different species of pollutants and atmospheric conditions and weight the importance of these data in the calculation of a single summative statistic. These metrics are useful in that they can quickly communicate a great deal of information in a form more easily understood by the public than raw air quality data [87]. Consolidated metrics can also change behavior in ways that minimize the impacts of air pollution, as demonstrated by AQHI-informed decisions to limit personal exposure [87].



The Environmental Performance Index (EPI) is an example of a consolidated metric that can be used in service of sustainability. The EPI is the a function of indicators such as urban particulates,  $\text{SO}_x$  and  $\text{NO}_x$  emissions, water stress, habitat protection, pesticide regulation, agricultural water intensity, and GHG emission per capita [131]. It has been applied to rank countries in terms of the state and treatment of their environment, and it can help countries define goals and track progress. It was created in response to a feeling that the environmental goals of the United Nations Millennium Development Goals were poorly defined and that progress toward these goals was difficult to quantify [131]. A remedy to these complaints is exactly what consolidated metrics and monitoring designed with these metrics in mind can provide. How well the EPI can describe a country's environmental status is limited by the data available [131]. The EPI can only incorporate data for which there is already a strong history and temporal trend, making it difficult to represent recently recognized sustainability issues. Nevertheless, the EPI provides countries with an evaluation as to where they stand relative to their peers, a method for tracking progress, and a databased framework to guide decision making.

The discipline of life-cycle assessment (LCA) contains many examples of consolidated metrics, similar to the EPI. LCA method and metrics can be used in attempts to quantify sustainability [133]. Eco-Indicator 99, for example, quantifies the impacts of activities and pollutant releases; translates these impacts into three metrics related to health, ecosystem damages, and depletion of resources; and combines these three metrics into a single indicator [134]. The weighting systems that allow LCA methods such as these to combine disparate data into a single number would encourage the quantification of the relative impacts of various air pollutants. Were that accomplished, data on different pollutants, pollutants with very different behavior and effects, could be compared on the same scale.

Designing air pollution monitoring and monitoring campaigns to serve consolidated sustainability metrics means not only providing data for metrics like the EPI and Eco-indicator 99, but also developing novel data sets to strengthen such metrics. Novel data sets would make it possible for such a metric to be

influenced by sustainability issues for which sufficient data do not yet exist. Collecting data so that loss of resilience, for example, could be incorporated into a metric like the EPI would encourage nations to take action on previously ignored sustainability issues and provide governments with the information necessary to do so. It must be borne in mind, however, that there are many pitfalls in the design of consolidated metrics [132], and only a correctly designed metric will lead to sound decision making.

A standardized set of data for use with widely accepted consolidated metrics would clarify the design goals of monitoring campaigns, which is important to the cost-efficiency of monitoring [13]. Information derived would be more similar from campaign to campaign, allowing monitoring groups to more easily share data. Further development of consolidated metrics in the direction of sustainability and the design of monitoring efforts to serve this development would allow the air pollution monitoring community to come to a consensus on the types of data to be counted in the metrics and the weighting of these data.

**Development of Satellite Observations** Satellite observations can be used in combination with models and in situ measurements to reduce statistical uncertainty in estimates [85]. Recent work by van Donkelaar et al. [86] has made possible the reliable estimation of long-term average  $\text{PM}_{2.5}$  concentration based on satellite-derived data. Aerosol Optical Depth (AOD) measurements made by the MODIS (Moderate Resolution Imaging Spectroradiometer) and MISR (Multiangle Imaging Spectrometer) instruments on the Terra satellite were combined with estimates of aerosol distribution from the Goddard Earth Observing System chemical transport model (GEOS-chem). Previously, variations in meteorology, aerosol composition, and distribution have prevented accurate estimates by satellite observation beyond small regions, but with the GEOS-Chem model, reliable estimates were obtained across all midlatitude areas of the globe. The concentration estimates with  $10 \text{ km}^2$  spatial resolution were obtained.

Many epidemiological studies of  $\text{PM}_{2.5}$  exposure are limited by a dearth of monitoring data, and the improved satellite estimates from the van Donkelaar et al. 2010 work [86], can now provide some of these

needed data [86]. The lack of data is more pronounced in developing countries, and these countries often have pronounced  $\text{PM}_{2.5}$  management problems, making these estimates particularly useful. van Donkelaar et al. [86] also suggest that increased satellite coverage could further improve the identification of temporal or spatiotemporal trends. In contrast, other recent attempts to estimate  $\text{PM}_{2.5}$  concentrations using satellite-observed AOD with fine spatial resolution determined that AOD is a poor tool for assessing spatial  $\text{PM}_{2.5}$  patterns [135], supporting previous suggestions that current satellite spatial resolution is not sufficient for the intra-urban scale [88]. AOD was also found to contribute little additional information when incorporated into a predictive model that was already accounting for land use, emissions sources, meteorology, and regional variability [135]. Thus, the most effective role of satellite data is still being established, but this tool has the potential to become a valuable and routine part of air pollution monitoring, providing data where there otherwise may be few and improving data where routine rigorous monitoring is taking place.

Martin [88] reviews the many recent improvements in the application of satellite observations to assessments of air quality but includes a number of limitations in obtaining accurate pollutant concentrations. Limitations included the need for more ground-level  $\text{NO}_2$  measurement to validate satellite-observed  $\text{NO}_2$  data and difficulty in obtaining high-quality  $\text{O}_3$  concentration estimates. Martin's work asserts that in terms of instrumentation, air quality monitoring via the current satellites is at a peak but that substantial progress could be made with a new generation of instruments, with more satellites in a variety of orbits (as opposed to all satellites being in sun-synchronous orbit as is currently the case). Specific instrument improvements include a minimization of pixel size, which will reduce the number of pixels obscured by clouds. Work that can be accomplished without additional satellites and a new generation of instrumentation includes the improvement of models critical to the interpretation of satellite observations. Better estimations from these models will improve information derived [88]. If the suggested developments can be made, satellite observations will move closer to delivering spatial and temporal data on a scale that better relates to ground-level monitoring.

## Design Methodologies for Monitoring Networks

All monitoring techniques, from measurement of ambient pollution levels to computer modeling, rely on some assumptions and estimates in the assessment of pollutant concentrations, exposure to pollution, risk, and damages from pollution. Sophisticated design methodologies for monitoring networks and campaigns have been developed to ensure that the uncertainties are minimized. These methodologies help to ensure that the most useful information can be extracted from what are often limited data.

Established design methodologies include kriging, the use of indices of multivariate effectiveness [136], and the location-allocation method [137]. Each method requires an established area to be monitored and attempts to optimally cover that area with a given number of monitoring stations. Kriging involves placing monitoring stations such that mean square prediction error is minimized [137]. The method employing indices of multivariate effectiveness is based on Shannon's [138] information theory, as used by Silva and Quiroz [136] which tests various configurations of monitors to optimize the information content of data collected. The location-allocation method begins with the creation of a "demand surface," a surface representing the area to be studied, with higher demand areas represented by peaks in the surface. Values on the demand surface are increased by variability in pollution concentrations and by proximity to sensitive populations [137]. To inform the demand surface, an initial pollution surface (representing the distribution of pollution over an area) must exist based on previous or ongoing monitoring. Once the demand surface is finalized, monitoring stations can be placed at the highest demand locations in the area.

Design methodologies for monitoring networks and sampling campaigns are a relatively new development, with many previous networks and campaigns set up in a subjective manner. However, past campaigns and networks lacking rigorous design principles have been effective in that they led to the first generations of information and significant discoveries relating to air pollution (and thereby led much of the effective legislation on air pollution). Further, these campaigns provided information that is necessary for the implementation of some of the newer design methodologies such as the location-allocation methodology,

and they have brought the state of air pollution science to the point where monitoring system design methodologies can be created. However, now that design methodologies have been established, their routine employment can improve the information content of monitoring data and the cost-effectiveness of monitoring efforts.

### Overcoming Oversimplifications in Monitoring

Often, the practicalities of air pollution monitoring necessitate compromises in aspects such as increased time resolution, assumptions about population behavior, and assumptions of relative uniformity over areas. As methods become more sophisticated, some of these assumptions can be eliminated.

Conventional studies usually assume that the basis of exposure to air pollution is the concentration in a given location, often outdoors. However, people move through various exposure situations in their daily lives, from indoors to “hotspot” intersections and through areas of differing land use. A dynamic activity-based population model has been suggested to develop a more realistic sense of actual individual exposure [139]. Use of this model captures exposure resulting from people’s travel in and out of urban areas, rather than assuming a constant number of people within cities, as has been the case in conventional studies [139]. Beckx et al. [139] use of this model showed that conventional static modeling of exposure is underestimating population exposure to air pollution.

Some very small areas in cities, such as high traffic intersections, can have much higher pollution concentrations than the bulk concentrations in the wider area being monitored through conventional means. Gokhale and Khare [141] discuss a model for determining the likelihood of air pollution concentration exceedances over norms at intersections and other areas, where exceedances are likely. The model provides the opportunity to probabilistically manage air pollution on a very fine spatial scale by designing for a minimization of exceedances. Models such as those of Beckx et al. [139], Burke et al. [140] and Gokhale and Khare [141] can provide means for addressing some of the simplifying assumptions in air pollution monitoring. Since their approaches primarily involve modeling, they need not increase the cost of air pollution management excessively.

### Monitoring More Effectively: Conclusion

There are a number of changes to air pollution monitoring suggested in the literature and, in many cases, already under way. These changes, the development of multipollutant science, monitoring for consolidated metrics, improvements to satellite-based monitoring, using monitoring network design methodologies, and dealing with routine oversimplifications will improve the operation of the monitoring community and the quality of results. Information, as a common first step toward sustainability and a necessary support for sustainability policy, will help the air pollution community serve sustainability. Hopefully, these changes will be only the first set of changes in a more fundamental shift toward sustainability.

### Conclusion

Air pollution is often seen as a negative but unavoidable consequence of human activity. While air pollutants are well classified and there are a number of commonly recognized problems they create, including negative impacts on human health, environmental damage, and economic losses, the establishments designed to mitigate these problems often have a mandate to also pursue sustainability.

The recommendations made herein regarding a shift toward sustainability are largely complimentary. Pursuing one of the goals described would serve other sustainability goals. That complementarity may be one of the most convincing arguments that these goals actually will promote sustainability. While sustainability is difficult to define, it must connote that solutions to one problem do not exacerbate other problems, but contribute to other solutions.

Holism, then, must be an aspect of sustainability, and the lack of holism in the current air pollution monitoring community may underlie its current shortcomings. Not only is there a need for more holistic viewpoints encompassing ideas from other intellectual disciplines, but there is a need for air pollution mitigation to go beyond promoting primarily technology-based solutions to only the immediate impacts. Solutions offered today do not address all aspects of the air pollution problem.

To a very large extent, the lack of holism is not a function of a failing of the air pollution monitoring community, but a result of that community having strategically chosen battles in the past. Dockery [43]

shows how hard advocates of air pollution mitigation were forced to struggle for any political action on air pollution in the early days of air pollution science, and how the battles were won by presenting irrefutable, solely technical and solely scientific evidence proving the very immediate and anthropocentric effects of air pollution. It is only now that less obvious, though arguably more dire, effects on humans, the systems they rely on for sustenance, and on nonhuman stakeholders are gaining more than a slight foothold within the political agenda.

Dockery [43] also shows the great extent to which the burden of proof was on those trying to defend people and ecosystems from the damages of pollution, and not on the polluters. This burden is still in place, and problems must be proven in terms polluters cannot ignore, such as dollar figures and mortality. To a large extent, evidence will need to continue to be in these terms, and monitoring must be expanded in volume, scope, and abilities to continue to prove the damages from air pollution in the normally accepted terms. However, the body of evidence will be more compelling and complete if the terms are expanded to include goals discussed in this article. The terms have been set as a reaction to refusals of those challenging the need for, or priority of, air pollution management, and reactionary strategies do not promote sustainability. If monitoring continues in this way, sustainability will remain elusive. If monitoring is performed to serve sustainability, sustainability will be served.

## Bibliography

### Primary Literature

1. US EPA (2010) What are the six common air pollutants? <http://www.epa.gov/air/urbanair/>. Accessed 24 Oct 2010
2. Environment Canada (2010) Criteria air contaminants and related pollutants. <http://www.ec.gc.ca/air/default.asp?lang=En&n=7C43740B-1>. Accessed 24 Oct 2010
3. Environment Canada (2010) Background: persistent organic pollutants. <http://www.ec.gc.ca/lcpe-cepa/default.asp?lang=En&n=135D347F-1>. Accessed 9 Dec 2010
4. Harrison RM (2001) Air pollution: sources, concentrations and measurements. In: Harrison RM (ed) *Pollution: causes, sources and control*, 4th edn. Royal Society of Chemistry, Cambridge, pp 169–193
5. US EPA (2008) Technology transfer network air toxics website. <http://www.epa.gov/ttn/atw/188polls.html>. Accessed 19 Oct 2010
6. Fenger J (1999) Urban air quality. *Atmos Environ* 33:4877–4900
7. Sillman S, Logan JA, Wofsy SC (1990) The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes. *J Geophys Res* 95:1837–1851
8. Mølhave L, Nielson GD (1992) Interpretation and limitations of the concept “total volatile organic compounds” (TVOCs) as an indicator of human responses to exposures of volatile organic compounds (VOC) in indoor air. *Indoor Air* 2:65–77
9. Wang DK, Austin CC (2006) Determination of complex mixtures of volatile organic compounds in ambient air: an overview. *Anal Bioanal Chem* 386:1089–1098
10. Kesselmeier J, Staudt M (1999) Biogenic volatile organic compounds (VOC): an overview on emission, physiology and ecology. *J Atmos Chem* 33:23–88
11. Geddes JA, Murphy JG, Wang DK (2009) Long term changes in nitrogen oxides and volatile organic compounds in Toronto and the challenges facing local ozone control. *Atmos Environ* 43:3407–3415
12. Brook RD, Rajagopalan S, Pope CA III, Brook JR, Bhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker RV, Mittleman MA, Peters A, Siscovick D, Smith SC Jr, Whitsel L, Kaufman JD (2010) Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American heart association. *Circulation* 121(21):2331–2378
13. Craig L, Brook JR, Chiotti Q, Croes B, Gower S, Hedley A, Krewski D, Krupnick A, Krzyzanowski M, Moran MD, Pennel W, Samet JM, Schneider J, Shortreed J, Williams M (2008) Air pollution and public health: a guidance document for risk managers. *J Toxicol Environ Health A* 71:588–698
14. Haywood JM, Shine KP (1995) The effect of anthropogenic sulfate and soot aerosol on the clear sky planetary radiation budget. *Geophys Res Lett* 22:603–606
15. Schecter A, Birnbaum L, Ryan JJ, Constable JD (2006) Dioxins: an overview. *Environ Res* 101:419–428
16. Fleischer O, Wichmann H, Lorenz W (1999) Release of polychlorinated dibenzo-p-dioxins and dibenzofurans by setting off fireworks. *Chemosphere* 39:925–932
17. Kampa M, Castanas E (2008) Human health effects of air pollution. *Environ Pollut* 151:362–367
18. WHO (World Health Organization) (1977) Environmental health criteria 3: lead. WHO, Geneva
19. US EPA (Environmental Protection Agency) (1997) The benefits and costs of the clean air act, 1970 to 1990. US EPA, Washington, DC
20. Crathorne B, Rees YJ, France S (2001) Chemical pollution of the aquatic environment by priority pollutants and its control. In: Harrison RM (ed) *Pollution: causes, sources and control*, 4th edn. Royal Society of Chemistry, Cambridge, pp 1–31
21. Gillett NP, Arora VK, Zickfeld K, Marshall SJ, Marryfield WJ (2011) Ongoing climate change following a complete cessation of carbon dioxide emissions. *Geoscience, Nature*. doi:10.1038/NGE01047
22. European Commission (1999) Directorate General, environment, nuclear safety and civil protection. EU Focus on Clean Air, Germany. [http://ec.europa.eu/environment/air/pdf/clean\\_air.pdf](http://ec.europa.eu/environment/air/pdf/clean_air.pdf)

23. Environment Canada (2010) Impacts of air pollution. <http://ec.gc.ca/indicateurs-indicators/default.asp?lang=en&n=D189C09D-1>. Accessed 17 Jan 2011
24. Yang W, Omaye ST (2009) Air pollutants, oxidative stress and human health. *Mutat Res* 674:45–54
25. Saxon A, Diaz-Sanchez D (2005) Air pollution and allergy: you are what you breathe. *Nat Immunol* 6:223–226
26. Mortimer KM, Neas LM, Dockery DW, Redline S, Tager IB (2002) The effect of air pollution on inner-city children with asthma. *Eur Respir J* 19:699–705
27. Lee SC, Lam S, Fai HK (2001) Characterization of VOCs, ozone, and PM<sub>10</sub> emissions from office equipment in an environmental chamber. *Build Environ* 36:837–842
28. Jerret M, Burnett RT, Pope A III, Ito K, Thurston G, Krewski D, Shi Y, Calle E, Thun M (1995) Long-term ozone exposure and mortality. *N Engl J Med* 360:1085–1095
29. Schwartz J, Morris R (1995) Air pollution and hospital admissions for cardiovascular disease in Detroit, Michigan. *Am J Epidemiol* 142:23–35
30. Beeson WL, Abbey DE, Knutsen SF (1998) Long-term concentrations of ambient air pollutants and incident lung cancer in California adults: results from the AHSMOG study. *Environ Health Perspect* 106:813–823
31. Woodruff TJ, Grillo J, Schoendorf KC (1997) The relationship between selected causes of postneonatal infant mortality and particulate air pollution in the United States. *Environ Health Perspect* 105:608–612
32. Pope CA III, Dockery DW (2006) Health effects of fine particulate air pollution: lines that connect. *J Air Waste Manage Assoc* 56:709–742
33. Pope CA III, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath CW Jr (1995) Particulate air pollution as a predictor of mortality in a prospective study of US adults. *Am J Respir Crit Care Med* 151:669–674
34. Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD (2002) Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J Am Med Assoc* 287:1132–1141
35. Stölzel M, Breitner S, Cyrys J, Pitz M, Wölke G, Kreyling W, Heinrich J, Wichmann H-E, Peters A (2007) Daily mortality and particulate matter in different size classes in Erfurt, Germany. *J Expo Sci Environ Epidemiol* 17:458–467
36. Delfino RJ, Sioutas C, Malik S (2005) Potential role of ultrafine particles in associations between airborne particle mass and cardiovascular health. *Environ Health Perspect* 113:934–946
37. Li N, Sioutas C, Cho A, Schmitz D, Misra C, Sempf J, Wang M, Oberley T, Froines J, Nel A (2003) Ultrafine particulate pollutants induce oxidative stress and mitochondrial damage. *Environ Health Perspect* 111:455–460
38. Vinzents PS, Möller P, Sørensen M, Knudsen LE, Hertel O, Jensen FP, Schibye B, Loft S (2005) Personal exposure to ultrafine particles and oxidative DNA damage. *Environ Health Perspect* 113:1485–1490
39. Knol A, de Hartog JJ, Boogaard H, Slottje P, van der Sluijs JP, Lebrecht E, Cassee FR, Wardekker A, Ayres JG, Brom PJ, Brunekreef B, Donaldson K, Forastiere F, Holgate ST, Kreyling WG, Nemery B, Pekkanen J, Stone V, Wichmann H-E, Hoek G (2009) Expert elicitation on ultrafine particles: likelihood of health effects and causal pathways. Part Fibre Toxicol 6:19. doi:10.1186/1743-8977-6-19
40. Finkelstein MM, Jerrett M (2007) A study of the relationships between Parkinson's disease and markers of traffic-derived and environmental manganese air pollution in two Canadian cities. *Environ Res* 104:420–432
41. Ewan KB, Pamphlett R (1996) Increased inorganic mercury in spinal motor neurons following chelating agents. *Neurotoxicology* 17:343–349
42. WHO (World Health Organization) (2010) Urbanization and Health. *Bull World Health Organ* 88:245–246
43. Dockery DW (2009) Health effects of particulate air pollution. *Ann Epidemiol* 19:257–263
44. Pope CA, Ezzati M, Dockery DW (2009) Fine-particulate air pollution and life expectancy in the United States. *N Engl J Med* 360:376–386
45. Likens GE, Driscoll CT, Buso DC, Mitchell MJ, Lovett GM, Bailey TG, Reiners WA, Alewell C (2002) The biogeochemistry of sulfur at Hubbard Brook. *Biogeochemistry* 60:235–315
46. Likens GE, Driscoll CT, Buso DC (1996) Long-term effects of acid rain: response and recovery of a forest eco-system. *Science* 272:244–246
47. Tomlinson GH (2003) Acidic deposition, nutrient leaching and forest growth. *Biogeochemistry* 65:51–81
48. Rechcigl JE, Sparks DL (1985) Effects of acid rain on the soil environment: a review. *Commun Soil Sci Plant Anal* 16:653–680
49. Koricheva J, Roy S, Vranjic JA, Haukioja E, Hughes PR, Hanninen O (1997) Antioxidant responses to simulated acid rain and heavy metal deposition in birch seedlings. *Environ Pollut* 95:249–258
50. Kratky BA, Fukunaga ET, Hylin JW, Nakano RT (1974) Volcanic air pollution: deleterious effects on tomatoes. *J Environ Qual* 3:138–140
51. Lazarus BE, Schaberg PG, Hawley G, DeHayes DH (2006) Landscape-scale spatial patterns of winter injury to red spruce foliage in a year of heavy region-wide injury. *Can J For Res* 36:142–152
52. Haines TA (1981) Acidic precipitation and its consequences for aquatic ecosystems: a review. *Trans Am Fish Soc* 110:669–707
53. Yi H, Hao J, Tang X (2006) Atmospheric environmental protection in China: current status, developmental trend and research emphasis. *Energy Policy* 35:907–915
54. Grantz DA, Garner JHB, Johnson DW (2003) Ecological effects of particulate matter. *Environ Int* 29:213–239
55. WHO (World Health Organization) (1991) Environmental health criteria 118: inorganic mercury. World Health Organization, Geneva
56. Kelting DL, Burger JA, Edwards GS (1995) The effects of ozone on the root dynamics of seedlings and mature red oaks (*quercus rubra* L.). *For Ecol Manage* 79:197–206
57. Lefohn AS, Jackson W, Shadwick DS, Knudsen HP (1997) Effect of surface ozone exposures on vegetation



- grown in the southern Appalachian Mountains: identification of possible areas of concern. *Atmos Environ* 31: 1695–1708
58. Laurence JA (1998) Ecological effects of ozone: integrating exposure and response with ecosystem dynamics and function. *Environ Sci Policy* 1:179–184
  59. Ren W, Tian H, Liu M, Zhang C, Chen G, Pan S, Felzer B, Xu X (2007) Effects of tropospheric ozone pollution on net primary productivity and carbon storage in terrestrial ecosystems of China. *J Geophys Res* 112:D22S09. doi:10.1029/2007JD008521
  60. Zhang Y, Zhu X, Slanina S, Shao M, Zeng L, Hu M, Bergin M, Salmon L (2004) Aerosol pollution in some Chinese cities. *Pure Appl Chem* 76:1227–1239
  61. OMA (Ontario Medical Association) (2005) The illness costs of air pollution: 2005–2026 health & economic damage estimates. OMA, Toronto
  62. Krupnick A, Alberini A, Cropper M, Simon N, O'Brien B, Goeree R, Heintzelman M (2002) Age, health and the willingness to pay for mortality risk reductions: a contingent valuation survey of Ontario residents. *J Risk Uncertain* 24:161–186
  63. Mayerhofer P, Weltschew M, Trukenmuller A, Friedrich R (1995) A methodology for the economic assessment of material damage caused by SO<sub>2</sub> and NO<sub>x</sub> emissions in Europe. *Water Air Soil Pollut* 85:2687–2692
  64. Delucchi MA (2000) Environmental externalities of motor-vehicle use in the US. *J Transp Econ Policy* 34:135–168
  65. Metallo MC, Poli AA, Diana M, Persia F, Cirillo MC (1995) Air pollution loads on historical monuments: an air quality model application to the marble arch of Titus in Rome. *Sci Total Environ* 171:163–172
  66. Murphy JJ, Delucchi MA, McCubbin DR, Kim HJ (1999) The cost of crop damage caused by ozone air pollution from motor vehicles. *J Environ Manage* 55:273–289
  67. Skärby L, Selldén G (1984) The effects of ozone on crops and forests. *Ambio* 13:68–72
  68. USDA, (United States Department of Agriculture) (2008) Budget summary and annual performance plan. USDA, Washington, DC
  69. Rasker R (1994) A new look at old vistas: the economic role of environmental quality in western public lands. *Univ Colo Law Rev* 65:369–399
  70. Gilbertson T, Reyes O (2009) Carbon trading: how it works and why it fails. Dag Hammarskjöld, Uppsala
  71. Solomon S (1999) Stratospheric ozone depletion: a review of concepts and history. *Rev Geophys* 37:275–316
  72. Weatherhead EC, Anderson SB (2006) The search for signs of recovery of the ozone layer. *Nature* 441(7089):39–45. doi:10.1038/nature04746
  73. Swart R, Amann M, Raes F, Tuinstra W (2004) A good climate for clean air: linkages between climate change and air pollution. *Clim Change* 66:263–269
  74. Harvey LD (2010) Energy and the new reality 1 – energy efficiency and the demand for energy. Earthscan, London. ISBN 10: 1849710724
  75. Environment Canada (2009) About the air quality health index. <http://www.ec.gc.ca/cas-aqhi/default.asp?Lang=En&n=065BE995-1>. Accessed 13 Mar 2010
  76. Wetherbee GA, Shaw MJ, Latysh NE, Lehmann MB, Rothert JE (2010) Comparison of precipitation chemistry measurements obtained by the Canadian air and precipitation monitoring network and national atmospheric deposition program for the period 1995–2004. *Environ Monit Assess* 164:111–132
  77. Acid Deposition Monitoring Network in East Asia (EANET) (2000) Technical manual for wet deposition monitoring in East Asia. <http://www.eanet.cc/product/techwet.pdf>
  78. Thepanondh S, Ayers GP, Hooper MA (2005) Analysis of precipitation chemistry in northern Thailand. *Clean Air Environ Qual* 39:43–47
  79. Jankhe JA (2000) Continuous emissions monitoring, 2nd edn. Wiley, New York. ISBN 0471292273
  80. Environment Agency (UK) (2008) Technical guidance note (monitoring) M20 – quality assurance of continuous emission monitoring systems – application of BS EN 14181 and BS EN 13284-2. <http://publications.environment-agency.gov.uk/pdf/GEHO1105BJXZ-e-e.pdf>
  81. Environment Canada (2005) Locomotive emissions monitoring program 2004. [http://www.railcan.ca/documents/publications/1118/2006\\_04\\_12\\_LemReport2004\\_en.pdf](http://www.railcan.ca/documents/publications/1118/2006_04_12_LemReport2004_en.pdf)
  82. Cooper DA, Ekström M (2005) Applicability of the PEMS technique for simplified NO<sub>x</sub> monitoring on board ships. *Atmos Environ* 39:127–137
  83. Williams ML (2001) Atmospheric dispersal of pollutants and the modeling of air pollution. In: Harrison RM (ed) *Pollution: causes, sources and control*, 4th edn. Royal Society of Chemistry, Cambridge, pp 246–267
  84. Miffre A, Chacra MA, Geffroy S, Rairoux P, Soulhac L, Perkins RJ, Frejafon E (2010) Aerosol load study in urban area by Lidar and numerical model. *Atmos Environ* 44:1152–1161
  85. Sarigiannis DA, Soulakellis NA, Sifakis NI (2004) Information fusion for computational assessment of air quality and health effects. *Photogramm Eng Remote Sens* 70: 235–245
  86. van Donkelaar A, Martin RV, Brauer M, Kahn R, Levy R, Verduzco C, Villeneuve PJ (2010) Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. *Environ Health Perspect* 118(6):847–855
  87. Brook JR, Moran MD, Pennell W, Craig L (2010) Elements of air quality management: atmospheric science tools for developing effective policy. In: Gurjar BR, Molina LT, Ojha CSP (eds) *Air pollution: health and environmental impacts*. Taylor & Francis, Boca Raton, pp 363–418
  88. Martin RV (2008) Satellite remote sensing of surface air quality. *Atmos Environ* 42:7823–7843
  89. Boersma KF, Eskes HJ, Veefkind JP, Brinksma EJ, van der A RJ, Sneep M, van den Oord GHJ, Levelt PF, Stammes P, Gleason JF, Bucscl EJ (2007) Near-real retrieval of tropospheric NO<sub>2</sub> from OMI. *Atmos Chem Phys* 7:2013–2128

90. Bachmann J (2007) Will the circle be unbroken: a history of the U.S. national ambient air quality standards. *J Air Waste Manage Assoc* 57:652–697
91. Chow JC, Watson JG (2008) New directions: beyond compliance air quality measurements. *Atmos Environ* 42:5166–5168
92. European Commission (2008) Air quality standards. <http://ec.europa.eu/environment/air/quality/standards.htm>. Accessed 19 Jan 2010
93. US EPA (Environmental Protection Agency) (2010) National ambient air quality standards (NAAQS). <http://www.epa.gov/air/criteria.html>. Accessed 19 Jan 2010
94. European Commission (2008) Commission regulation (EC) no. 692/2008, O.J. L 199/1
95. US EPA (Environmental Protection Agency) (2008) National air quality: status and trends through 2007. US EPA Office of Air Quality Planning and Standards, Air Quality Assessment Division, Research Triangle Park
96. Environment Canada (2005) Acid rain and the facts. <http://www.ec.gc.ca/AcidRain/acidfact.html>. Accessed 25 Jan 2010
97. Wheeler D (2001) Racing to the bottom? Foreign investment and air pollution in developing countries. *J Environ Dev* 10:225–245
98. Fredriksson PG, Gaston N (2007) Environmental governance in federal systems: the effects of capital competition and lobby groups. *Econ Inq* 38:501–514
99. Hidy GM, Brook JR, Demerjian KL, Molina LT, Pennell WT, Scheffe RD (eds) (2011) Technical challenges of multipollutant air quality management. Springer, Dordrecht. ISBN 9789400703032
100. Canada. Environment Canada and Health Canada (2011) Canadian smog science assessment (Unpublished)
101. Dominici F, Peng RD, Barr CD, Bell ML (2010) Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology* 21:187–194
102. Sundqvist G, Letell M, Lidskog R (2002) Science and policy in air pollution abatement strategies. *Environ Sci Policy* 5:147–156
103. Fayolle F, Vandecasteele J-P, Monot F (2001) Microbial degradation and fate in the environment of methyl *tert*-butyl ether and related fuel oxygenates. *Appl Microbiol Biotechnol* 56:339–349
104. Rass-Hansen J, Falsig H, Jørgensen B, Christensen C (2007) Bioethanol: fuel or feedstock? *J Chem Technol Biotechnol* 82:329–333
105. Statistics Canada (2007) Canadian environmental sustainability indicators: air quality indicators: data sources and methods. Statistics Canada, Ottawa
106. Brook JR, Demerjian KL, Hidy G, Molina L, Pennell W, Scheffe R (2009) New directions: results-oriented multi-pollutant air quality management. *Atmos Environ* 43:2091–2093
107. Law K (2010) Atmospheric chemistry: more ozone over North America. *Nature* 463:307–308
108. MacDonald DV, Hanley N, Moffat I (1999) Applying the concept of natural capital criticality to regional resource management. *Ecol Econ* 29:73–87
109. Folke C (2006) Resilience: the emergence of a perspective for social-ecological systems analyses. *Glob Environ Change* 16:253–267
110. Costanza R, d'Arge R, de Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill R, Paruelo J, Raskin RG, Sutton P, van den Belt M (1997) The value of the world's ecosystem services and natural capital. *Nature* 387:253–260
111. Pimm SL (1997) The value of everything. *Nature* 387:231–232
112. Daily GC, Söderqvist T, Aniyar S, Arrow K, Dasgupta P, Ehrlich PR, Folke C, Jansson A, Jansson B-O, Kautsky N, Levin S, Lubchenco J, Mären K-G, Simpson D, Starret D, Tilman D, Walker B (2000) The value of nature and the nature of value. *Science* 289:395–396
113. Engelbrecht H-J (2009) Natural capital, subjective well-being, and the new welfare economics of sustainability: some evidence from cross-country regions. *Ecol Econ* 69:380–388
114. Deutsch L, Folke C, Skanberg K (2003) The critical natural capital of ecosystem performance as insurance for human well-being. *Ecol Econ* 44:205–217
115. Rockstrom J, Steffen W, Noone K, Persson Å, Chapin FS III, Lambin EF, Menton TM, Scheffer M, Folke C, Schellnhuber HJ, Nykvist B, de Wit CA, Hughes T, van der Leeuw S, Rodhe H, Sörlin S, Snyder PK, Costanza R, Svedin U, Falkenmark M, Karlberg L, Corell RW, Fabry VJ, Hansen J, Walker B, Liverman D, Richardson K, Crutzen P, Foley JA (2009) A safe operating space for humanity. *Nature* 461:472–475
116. Folke C, Carpenter S, Elmqvist T, Gunderson L, Holling CS, Walker B (2002) Resilience and sustainable development: building adaptive capacity in a world of transformations. *Ambio* 31:437–440
117. O'Neill MS, Jerret M, Kawachie I, Levy JI, Cohen AJ, Gouveia N, Wilkinson P, Fletcher T, Cifuentes L, Schwartz J (2003) Health, wealth and air pollution: advancing theory and methods. *Environ Health Perspect* 111:1861–1870
118. Künzli N, Jerret M, Mack WJ, Beckerman B, LaBree L, Gilliland F, Thomas D, Peters J, Hodis HN (2005) Ambient air pollutant and atherosclerosis in Los Angeles. *Environ Health Perspect* 113:201–206
119. Mitchell G, Dorling D (2003) An environmental justice analysis of British air quality. *Environ Plann A* 35:909–929
120. MacKerron G, Mourato S (2009) Life satisfaction and air quality in London. *Ecol Econ* 68:1441–1453
121. Marshall JD, Brauer M, Frank LD (2009) Healthy neighbourhoods: walkability and air pollution. *Environ Health Perspect* 117:1752–1759
122. Dannenberg AL, Jackson RJ, Frumkin H, Schieber RA, Pratt M, Kochtitsky C, Tilson HH (2003) The impact of community design and land-use choices on public health: a scientific research agenda. *Am J Public Health* 93:1500–1508
123. Stone B Jr (2003) Air quality by design: harnessing the clean air act to manage metropolitan growth. *J Plann Educ Res* 23:177–190
124. Frank LD, Sallis JF, Conway TL, Chapman JE, Saelens BE, Bachman W (2009) Many pathways from land use to health. *J Am Plann Assoc* 72:75–87

125. Tang UW, Wang ZS (2007) Influences of urban form on traffic-induced noise and air pollution: results from a modelling system. *Environ Modell Softw* 22:1750–1764
126. Newman P, Beatley T, Boyer H (2008) *Resilient cities: responding to peak oil and climate change*. Island Press, Washington, DC. ISBN 9781597264990
127. Marshall JD, Brauer M, Frank LD (2009) Health neighbourhoods: walkability and air pollution. *Environ Health Perspect* 117:1752–1759
128. Royal Society (2006) Survey of factors affecting science communication by scientists and engineers. [http://royalsociety.org/uploadedFiles/Royal\\_Society\\_Content/Influencing\\_Policy/Themes\\_and\\_Projects/Themes/Governance/Final\\_Report\\_-\\_on\\_website\\_-\\_and\\_amended\\_by\\_SK\\_no\\_navigation.pdf](http://royalsociety.org/uploadedFiles/Royal_Society_Content/Influencing_Policy/Themes_and_Projects/Themes/Governance/Final_Report_-_on_website_-_and_amended_by_SK_no_navigation.pdf). Accessed 31 Mar 2010
129. Yearly S (2006) Bridging the science – policy divide in urban air-quality management: evaluating ways to make models more robust through public engagement. *Environ Plann C Gov Policy* 24:701–714
130. Pereira AG, Raes F, TdeS P, Rosa P, Broderson S, Jorgensen MS, Ferreira F, Querol X, Rea J (2009) Atmospheric composition change research: time to go post-normal? *Atmos Environ* 43:5423–5432
131. Sather ME, Varns JL, Mulik JD, Glen G, Smith L, Stallings C (2001) Passive ozone network of Dallas: a modeling opportunity with community involvement 2. *Environ Sci Technol* 35:4426–4435; Emerson J, Esty DC, Kim C, Srebotnjak T, Levy MA, Mara V, de Sheribinin A, Jaiteh M (2010) 2010 environmental performance index. Yale Center for Environmental Law & Policy. [http://epi.yale.edu/file\\_columns/0000/0157/epi2010\\_report.pdf](http://epi.yale.edu/file_columns/0000/0157/epi2010_report.pdf)
132. Swamee PK, Tyagi A (2007) Improved method for aggregation of water quality Subindices. *J Environ Eng* 133:220–225
133. Azapagic A (2006) Life cycle assessment as an environmental sustainability tool. In: Dewulf J, Van Langhove H (eds) *Renewables-based technology: sustainability assessment*. Wiley, Chichester, pp 87–110
134. Goedkoop M, Spriensma R (2001) *The Eco-indicator 99: a damage oriented method for life cycle assessment – methodology report*, 3rd edn. Pré Consultants, Amersfoort, The Netherlands
135. Paciorek CJ, Liu Y (2009) Limitations of remotely sensed aerosol as a spatial proxy for fine particulate matter. *Environ Health Perspect* 117:904–909
136. Silva C, Quiroz A (2003) Optimization of the atmospheric pollution monitoring network at Santiago de Chile. *Atmos Environ* 37:2337–2345
137. Kanaroglou PS, Jerret M, Morrison J, Beckerman B, Arain MA, Gilbert NL, Brook JR (2005) Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach. *Atmos Environ* 39:2399–2409
138. Shannon CE (1948) A mathematical theory of communication. *Mobile Comput Commun Rev* 5:3–55
139. Beckx C, Panis LI, Arentze T, Janssen D, Torfs R, Broekx S, Wets G (2009) A dynamic activity-based population modelling

approach to evaluate exposure to air pollution: methods and application to a Dutch urban area. *Environ Impact Assess Rev* 29:179–185

140. Burke JM, Zufall MJ, Ozkaynak H (2001) A population exposure model for particulate matter: case study results for PM<sub>2.5</sub> in Philadelphia, PA. *J Expo Anal Environ Epidemiol* 11:470–489
141. Gokhale S, Khare M (2007) A theoretical framework for the episodic-urban air quality management plan (e-UAQMP). *Atmos Environ* 41:7887–7894

## Books and Reviews

- Young TK (2005) *Population health: concepts and methods*, 2nd edn. Oxford University Press, New York. ISBN 0-19-515854-7

## Air Pollution Sources, Statistics and Health Effects, Introduction

ROY M. HARRISON

Division of Environmental Health and Risk Management, School of Geography, Earth and Environmental Sciences, The University of Birmingham, Edgbaston, Birmingham, UK

Air pollution has been a topical subject for many centuries. There are records of the nuisance effects of coal burning in London going back for hundreds of years, and the association of high air pollution levels with effects on human health has been recognized for almost as long. Urban pollution in the form of smogs arising from coal burning and industrial emissions, most notably in London, Donora (Pennsylvania), and the Meuse Valley in Belgium, was associated with marked increases in daily mortality and led to legislation which has, in the main, been very effective at improving air quality. Nonetheless, research over the last 2 decades has led to a change in paradigm in our understanding of the health effects of air pollutants. Whereas previously it was assumed that there were no-effect thresholds, i.e., concentrations below which adverse effects on health would not be observed, it is now recognized that for most toxic air pollutants, there is no recognizable no-effect threshold concentration, and effects are seen in the most sensitive members of the population at extremely low concentrations. The most influential studies of pollutants such as airborne

particulate matter have failed to reveal no-effect thresholds, and consequently, air quality management now seeks to limit rather than wholly eliminate adverse effects on the population. The pollutant generally accepted as having the greatest public health impact is particulate matter. Both short-term (acute) and long-term (chronic) exposures have effects upon health. The best quantified information relates to mortality and hospital admissions. Exposure to particulate matter in the  $PM_{10}$  size range is associated with an increase of around 0.6% in mortality for a 24-h increase of  $10 \mu g m^{-3}$ . The impacts of chronic exposure are of greater public health significance with an increase of around 4% in mortality for a  $10 \mu g m^{-3}$  increase in particles in the  $PM_{2.5}$  size range. The impact on life expectancy is a reduction of between a few months and several years across Europe depending upon the local  $PM_{2.5}$  concentrations. Adverse effects on human health have been the main driver of scientific research on the emission, atmospheric behavior, and health effects of toxic pollutants referred to as “locally acting” pollutants. Strict *air quality guidelines and standards*, including limit values, have been adopted in most countries in order to limit the adverse effects of exposure to a range of pollutants, most typically including particulate matter, sulfur dioxide, nitrogen dioxide, ozone, lead, and carcinogenic pollutants such as benzene and the polycyclic aromatic hydrocarbons.

The other main driver of science in the air pollution field has been the recognition that air pollutants have an effect upon global climate and that some of the locally acting air pollutants also have the capacity to influence climate locally, regionally, and globally; for example, black carbon from combustion sources has much the same effect as a greenhouse gas, although the precise mechanisms are different. Airborne particles also affect radiative transfer in the atmosphere and provide one of the largest uncertainties in estimating the anthropogenic influences upon climate change. White particles such as ammonium sulfate are reflective and have a net cooling effect by reflecting incoming solar radiation back to space. Water-soluble particles can act as cloud condensation nuclei affecting the albedo (reflectivity) of clouds leading to a reduction in land surface warming.

The majority of the world's population now lives within urban areas which are also a focus for road

traffic activity and frequently also for power plants and industry. In this context, urban areas are often more polluted than the regional atmosphere surrounding them, and because of their high concentration of population, this is where a large part of human exposure to locally acting pollutants occurs. The entry [► Urban Air Quality: Sources and Concentrations](#) considers the sources of pollutants influencing urban areas, reviews the measured concentrations within cities internationally, and describes some of the approaches to air quality management. The determinants of urban concentrations of pollutants are the rates of emission of primary pollutants (those emitted directly from sources such as factory chimneys and engine exhaust), the meteorological processes that determine the mixing of the emissions within the atmosphere and the transport both horizontally and vertically away from the source, and the atmospheric chemistry, which can both cleanse the atmosphere of toxic substances but also lead to the formation of secondary pollutants, a term describing pollutants formed within the atmosphere itself. The meteorological processes leading to the vertical mixing of pollutants away from sources and the horizontal transport by the wind (termed advection) are described in the entry [► Urban Air Quality: Meteorological Processes](#). Computational models used to describe these processes are also included. The complementary entry [► Urban Atmospheric Composition Processes](#) describes the key chemical reactions which serve both to limit the concentrations of toxic pollutants and also to control concentrations of secondary pollutants such as ozone and secondary airborne particles which are very important in relation to both human health and regional pollution. Both photochemical and thermal reactions in the atmosphere can lead to formation of reactive free radical species which degrade organic compounds, sulfur dioxide, nitrogen dioxide, and many other toxic species. However, in doing so, they form secondary products, some of which can also have important consequences for health and climate. For example, oxidation of sulfur dioxide and nitrogen dioxide lead ultimately to the formation of ammonium sulfate and ammonium nitrate which in turn lead to increased concentrations of particulate matter and cloud condensation nuclei. As air pollutants are advected from urban areas, they are subject to atmospheric transport leading to impacts upon

► [regional air quality](#) which are important in relation both to human health effects and ecosystem effects described in the entry of this name. Transboundary air pollution, in which pollutants cross national boundaries is an important aspect of regional pollution. Chemistry-transport models, which simulate the processes involved in regional transport of pollution and are very important in policy development, are also described.

Most developed countries have achieved considerable success in improving air quality. This has depended upon an advanced approach to air quality management, which requires a number of key components dealt with in subsequent entries. One of the keys to the management of air pollution is high-quality monitoring of airborne concentrations. This is normally carried out with fixed ground-level monitors located so as to reflect outdoor exposures of the public. The entry ► [Air Pollution Monitoring and Sustainability](#) discusses the design of monitoring networks and how monitoring data can inform urban design and air quality management policy. In addition to traditional monitoring networks, the monitoring of source emissions and also regional remote sensing of atmospheric composition from space have an important role to play. In addition to monitoring, air quality management also depends upon ► [air quality guidelines and standards](#). The original intention of air quality guidelines was to provide no-effect concentrations which could be used as targets in air quality management. As explained above, many pollutants are now recognized as not having a demonstrable no-effect threshold, and consequently, air quality guidelines and standards are designed to limit adverse effects to a very low level, and the implementation of abatement policies is frequently based on cost-benefit analysis. One of the main sources of urban and regional pollution has been surface transportation and especially road traffic. More recently, there has been recognition of the impacts of shipping and railways on emissions and air quality, and the entry ► [Air Quality, Surface Transportation Impacts on](#) seeks to quantify the issues and place them in context. Abatement measures applied to road transport have been extremely effective in the developed world in limiting the harmful effects of pollutants emitted from vehicle exhausts, but much remains to be done to achieve full sustainability of the transport fleet. Aviation is also a significant source of pollutant

emissions, but because of the altitude at which most of the emissions occur, the effects on ground-level air quality are modest, and the main concern is over climate impacts arising from emissions from greenhouse gases and water vapor which are described in the entry ► [Aviation and Atmosphere](#). Release of long-lived pollutants at ground-level as well as aircraft emissions lead to ► [stratospheric pollution](#) with consequences for stratospheric ozone concentrations, penetration of UV radiation to ground-level, and for global climate.

Air pollutants comprise both gases and particles. The latter are tiny airborne liquid or solid droplets ranging from nanometers to tens or even hundreds of micrometers in size. In relation to locally acting pollution, they are typically referred to as particulate matter, but in the context of the global atmosphere, they are referred to as aerosol. The impacts of ► [aerosol in global atmosphere](#) of greatest concern are those upon climate. These are numerous and complex, and aerosol in the global atmosphere may have both warming and cooling effects upon climate. There are sometimes policy tensions between global and local pollution in that the abatement of some kinds of local pollution may worsen global warming problems. This entry deals with the sources and properties of aerosols in the global atmosphere.

Taken together, these contributions by eminent authors give a broad and rather comprehensive view of air pollution, extending from key urban sources and processes through the regional scale to the global, with close attention given to key sources such as road and air transportation.

---

## Air Quality Guidelines and Standards

ROGER O. MCCLELLAN

Toxicology and Health Risk Analysis, Albuquerque, NM, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Air Quality Management Framework

Context for Air Quality Impacts

The US Legal Framework for Air Quality Standards



National Ambient Air Quality Standards  
 US Process for Setting National Ambient Air Quality Standards  
 Paradigm Shift  
 WHO Process for Setting Air Quality Guidelines  
 Assessing Hazards and Risks  
 Sources of Information for Setting Standards and Guidelines  
 Particulate Matter as an Example  
 Ozone  
 Current Air Quality Standards and Guidelines  
 Future Directions  
 Bibliography

## Glossary

**Adverse health effect** Alteration in the health status of individuals or populations.

**Aerodynamic diameter** The diameter of a spherical particle of unity density with a terminal velocity equal to the particle being sampled.

**Aerosol** Relatively stable suspension of particles or droplets in a gaseous media.

**Air quality guidelines** AQG, established by World Health Organization to provide guidance to the public and government on levels of air pollutant that will improve air quality.

**Air quality monitoring** Measurement of the concentration of various pollutants in air for regulatory compliance or research purposes.

**Ambient air** Air found out of doors, which is a mixture of gases and suspended particles and droplets.

**Ambient concentration-response function** A mathematical description of the relationship between ambient air concentrations and associated morbidity or mortality indicators, usually expressed as increased risk for a defined increment of increased pollutant.

**Anthropogenic** Causal by humans, man-made.

**Background levels** The concentration of a pollutant in the air after accounting for all local or regional emissions.

**Biogenic** Produced by biological processes.

**Carbon monoxide** The chemical, CO, that is a product of incomplete combustion. It is regulated

as a Criteria Air Pollutant or Common Air Pollutant.

**Emissions** The quantities of pollutants emitted by a particular source or multiple sources.

**Excess risk** Risk associated with exposure to an agent over and above the baseline risk.

**Exposure** Taking materials into the body via inhalation, ingestion, or absorption through the skin.

**Guidelines** Guidance provided for use by the public or government officials.

**Hazardous** The property of having the potential to cause adverse health effects with sufficient duration and concentration of exposure.

**Lead** Pb, an element that occurs naturally and found many uses including as a paint additive and as tetraethyl Lead added to gasoline to enhance engine performance. Regulated as a Criteria Air Pollutant or Hazardous Air Pollutant.

**Mobile sources** Mobile sources, such as vehicles, of air pollution emissions.

**Morbidity** A statistical measure of number of fractions of indicators of illness.

**Mortality** A statistical measure of number or fraction of deaths.

**National Ambient Air Quality Standards** NAAQS, established under the US Clean Air Act to protect health and welfare.

**Nitrogen oxides** NO<sub>x</sub>, a mixture of NO and NO<sub>2</sub> produced in high-temperature combustion processes. Regulated as a Criteria Air Pollutant or Common Air Pollutant.

**Ozone** O<sub>3</sub>, a molecule with three atoms of oxygen. It is highly reactive with other chemicals and biological systems. It is regulated as a Criteria Air Pollutant or Common Air Pollutant.

**Particulate matter** Airborne solid material, unspecified as to chemical composition, regulated as a Criteria Pollutant or Common Air Pollutant.

**Particulate matter, 10 µm** PM<sub>10</sub>, airborne particulate material included in sampled air size fraction with aerodynamic size of 10 µm and less.

**Particulate matter, 2.5 µm** PM<sub>2.5</sub>, particulate material included in sampled air size fraction with an aerodynamic size of 2.5 µm or less.

**Primary standards or guidelines** Air Quality Standards or Guidelines for Criteria or Common Pollutants set to protect public health.

**Risk** The probability of an adverse health outcome associated with exposure to a hazardous material for a defined period of time and exposure level. Usually expressed as a level of adverse effect, morbidity, or mortality, over and above background.

**Risk assessment** The analytical process of evaluating exposure and the hazard potential of a material to estimate risk.

**Secondary standards or guidelines** Air Quality Standards or Guidelines for Criteria or Common Air Pollutants set to protect public welfare such as visibility or effects on plants.

**Standards** Regulations that are legally binding.

**Stationary sources** Fixed sources of air pollution emissions such as power plants.

**Sulfur dioxide** SO<sub>2</sub>, released from combustion of sulfur-containing fossil fuels. It is regulated as a Criteria Air Pollutant or Common Air Pollutant.

## Definition of the Subject

Air quality is the state of the atmosphere in which humankind lives and works. The air consists of a complex mixture of gases and suspended droplets and particles. Some of the gases, such as oxygen and carbon dioxide, within rather tight concentration limits, are essential for sustaining the life of all living things. Some of the gases and aerosol particles and droplets, when certain concentrations are exceeded and with sufficient duration of exposure, may have adverse effects on humans and other living things, including vegetation. Various governmental agencies around the world have developed air quality guidelines and standards for ambient or outdoor air quality. These guidelines or standards, when achieved or attained, are intended to be protective of human health and minimize the potential for other effects such as impacts on visibility or plants. Some government entities and statutes characterize the guidelines as goals to be attained, while other government entities and statutes treat them as rigid standards which are legally enforced. Guidelines or standards, when attained, are intended to improve the quality and sustainability of human life.

## Introduction

The quality of the air that people live in and breathe has no doubt been of concern to people since early in the evolution of humans. This concern, which must have preceded recorded history, likely coincided with man's use of purposeful combustion, fire, for cooking and heating and ultimately, for manufacturing and other activities, including transportation. There is abundant evidence that early humans found it convenient to live in caves both for shelter and safety. The poor quality of the air is apparent from the residual soot coating the interior of caves used as dwellings. Indeed, pictographs of the interior of the caves provide documentation of aspects of the life of those early cave dwellers. As community life evolved, people continued to live and use fire for both cooking and heating in poorly ventilated dwellings.

Domestication of both plants and animals played a critical role in the development of agriculture-based societies and the growth of communities. The agricultural base influenced the size of communities and their proximity to one another. The ability to transport essential foodstuffs by ship or overland by horses, camels, and oxen allowed cities to emerge. With larger communities came increased air pollution and the impact of an individual family's pollution on their neighbors and vice versa.

Coal apparently was first used as a fuel in the Far East. The introduction of coal as an energy source in the Western world followed Marco Polo's return to the west from his trip to the more advanced societies of the Far East. As early as the thirteenth century, there was concern over coal smoke and odor in London, and the Queen of England reportedly moved from London to Nottingham because of the insufferable smoke. In 1661, John Evelyn published *Fumifugium: Or the Inconvenience of the Air and Smoke of London Dissipated*, which drew attention to the problem for air pollution.

The industrial revolution in the early years was dependent on wood as a fuel, and therefore, industry was located near forests. This restriction served as a throttle on industrial development. The development of effective steam engines in the eighteenth century provided a major impetus for the use of coal as a fuel. With coal, the pace of industrial activity increased, and factories developed near sources of coal and along the

waterways that served to efficiently transport the coal. The smoke and ash from coal-fired boilers in factories, power plants, and locomotives increasingly became of concern. Interestingly, this concern about air pollution was stimulated by impairment of visibility and soiling as well as effects on health. In addition, the damage to vegetation from emissions produced in the smelting of sulfide ores gave clear evidence of yet another effect of air pollution. Air pollution problems were recognized in many industrial countries in the late nineteenth century, but the problem was probably nowhere as apparent as in London. The high levels of pollutants combined with the notorious London fog created a serious problem and is reputed to have given rise to the term *smog*, a contraction of smoke and fog.

The development of steam engines soon led to their use in sea-going vessels that initially depended heavily on coal as a fuel. Later, coal would be replaced by oil as a fuel. Emissions from ships did not attract major attention until the late twentieth century when concern developed for air emissions when the ships were near land and especially when in harbor.

With the discovery of electricity a revolution occurred in the generation and use of energy. More appropriately, it is the conversion of energy from one form to another, for example, from coal to thermal energy to electric energy that could be easily transmitted. It was now possible to generate electricity with coal as an energy source at one location and transport it over long distances for use at multiple locations. Over the course of the twentieth century the generation of electricity from the combustion of fossil fuels in large power plants continuously grew, first in heavily industrialized countries and, ultimately, around the world.

In the 1900s, a series of incidents drew attention to air pollution as a serious and even acute health problem. In December 1930, static atmospheric conditions in the Meuse Valley of Belgium resulted in pollution that caused over 60 deaths and hundreds of illnesses. In December 1952, smog in London resulted in about 4,000 deaths, principally among the infirm, the old, and those with respiratory diseases. The effects were not restricted to humans. Acute respiratory symptoms were also reported in cattle at the Smithfield Club livestock show. The effects may have been most severe in the heavy-weight prize cattle, who were well cared for and had their litter and excreta promptly removed.

Other cattle that were not as well cared for apparently fared better, perhaps because of some benefits of ammonia from the excreta neutralizing the effects of sulfur oxides and sulfuric acid. Another major smog episode in London during December 1962 resulted in 340 deaths.

In October 1948, a particularly calm and stable meteorology in Donora, Pennsylvania, about 30 km from Pittsburgh, PA, resulted in a 4-day buildup in levels of oxides and inorganic sulfates from the local steel and zinc smelters. More than 20 individuals died and 6,000 people developed respiratory problems. The official government report [57] on the Donora episode noted that the health effects were likely related to a combination of contaminants and identified sulfur dioxide as a chief culprit. The word smoke did not appear in the report. The attention previously focused on smoke was now broadening to include other airborne materials and, indeed, more broadly to air quality.

Despite early concerns about whether airborne materials could cause disease, relatively little research was conducted prior to World War I to improve our knowledge of the relationship between the air pollution and disease. The unfortunate use of poisonous gases in World War I provided a stimulus for developing a better understanding of how these specific agents produced disease. Unfortunately, the meager literature generated indicates that research in this area was not substantial or very long-lasting. In retrospect, the lack of research related to airborne toxicants is probably not surprising considering the generally low level of biomedical research activity in the first third of the twentieth century. The strong preventive medicine orientation that exists today toward diseases of both occupational and environmental origin had not yet developed.

Between World War I and World War II, concern for airborne materials impacting on human health centered primarily on occupational exposures. World War II brought with it renewed concern for purposeful exposures of people to toxic gases and biological agents. In both the United Kingdom and the United States, major research efforts were conducted to better understand the effects of these agents thereby providing improved approaches to defending against their use and effects. World War II also brought with it major

research efforts focusing on the potential toxic effects of uranium and newly discovered radionuclides such as plutonium and fission product related to the manufacture and use of nuclear weapons and, later, the development of nuclear reactors as energy sources to generate electrical power. These activities are noteworthy since many of the basic concepts of aerosol science and inhalation toxicology trace their roots to those research efforts.

The post-World War II era brought with it an increased awareness of air pollution arising from another source – motor vehicles. Perhaps nowhere was this more the case than in the Los Angeles basin with its marked increase in population, industry, and cars. As concern increased for the Los Angeles smog in the late 1940s and 1950s, the name “Bay of Smokers,” which was coined in 1542 by Juan Rodriguez Cabrillo for San Pedro Bay, took on a contemporary tone. Haagen-Smit [25] of the California Institute of Technology reported the scientific basis for photochemical smog – reactions among oxides of nitrogen and hydrocarbons from vehicle exhaust in the presence of sunlight produced ozone and other photochemical oxidants that were the key components of Los Angeles smog. It is noteworthy that his early research was motivated by understanding the effects of smog on citrus fruit production. Haagen-Smith’s findings were another key factor in broadening the concern for air quality from smoke to ozone and other air contaminants.

Post WWII then was a major and continuing increase in the use of airplanes both to transport people and goods. Indeed, transport of goods by aircraft is a central element of the global economy that emerged in the latter part of the twentieth century. As with ships, aircraft as sources of air pollution did not attract much attention in the late twentieth century when attention began to focus on air quality in the vicinity of major airports. In the case of some areas, such as the Los Angeles International Airport and a nearby harbor, concern has developed for the contribution of multiple sources: aircraft, cargo ships, locomotives, trucks, and passenger cars.

So far the discussion has centered on air pollution from combustion of hydrocarbons, from recently fixed carbon in wood and other plant materials to the fossil fuel energy sources, coal, oil, and natural gas, created in

earlier geologic eras. With each of these fuel sources hydrocarbons are combusted releasing thermal energy, CO, CO<sub>2</sub>, H<sub>2</sub>O, variable amounts of Nitrogen Oxides, and traces of other elements.

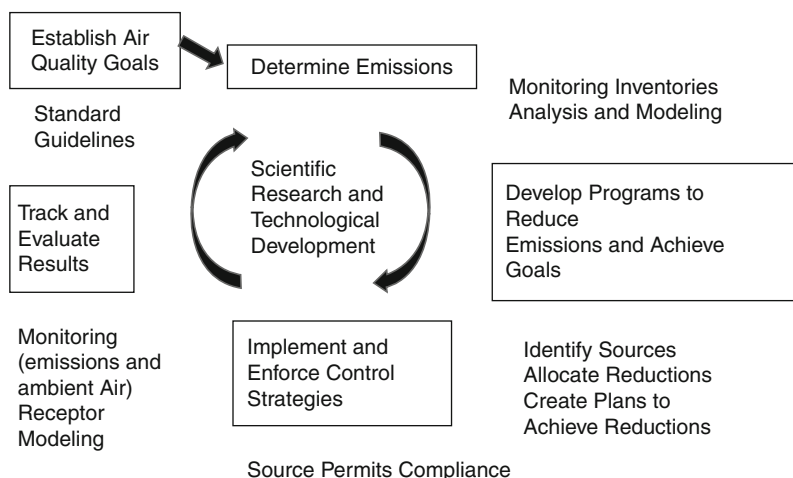
There are non-hydrocarbon-based energy systems that do not depend on combustion of hydrocarbons that deserve mention. This includes wind, hydro, nuclear, and solar power. Some of these such as wind and water have been in use for centuries to provide power. Nuclear power had its advent post-WWII. Solar power with its current focus on photovoltaic conversion is relatively new. All of these systems have the advantage of not creating airborne emissions related to combustion. However, each of these energy sources has its own disadvantages related to potential impacts on human health and ecosystems.

It is important to recognize that there are a multitude of other industrial and personal activities beyond combustion of hydrocarbons that result in air emissions that may be harmful to human health and ecosystems. These range from mining and extraction of minerals to processes producing specialized products such as paper, fabric, silicon chips, and chemicals to the manufacture of vehicles and other commercial goods. Agricultural activities also have unique airborne emissions. Each of these activities require variable amounts of energy and, thus, contribute to the potential for local emissions of combustion products or emissions at a remote location such as a central facility producing electrical power. In addition, each of these activities has airborne emissions unique to the processes being used.

The focus of this entry is on the pollutants that arise from multiple sources and, thus, are found everywhere. These common air pollutants are identified as criteria pollutants in the Clean Air Act, the principal statute governing air quality in the United States [14]. Some of these same pollutants have been identified by the World Health Organization (WHO) as common air pollutants [94].

## Air Quality Management Framework

Air quality guidelines and standards are a key element of air quality management as practiced around the world. Bachmann [7] provides an informative historical review of the evolution of air quality management in the United States with particular emphasis on the



**Air Quality Guidelines and Standards. Figure 1**  
Conceptual framework for air quality management

National Ambient Air Quality Standards (NAAQS). A conceptual framework for Air Quality Management is depicted in Fig. 1. Air quality goals, which may be expressed as air quality guidelines or standards, are the drivers for all the other elements of the framework. Emissions from various sources are considered within the context of the guidelines or standards. Are specific sources or kinds of sources significant contributors based on estimates of emissions? If so, what strategies can be used to reduce their emissions? Do source permits have a role in setting emission limits? Ambient monitoring can be used to provide an overall assessment of air.

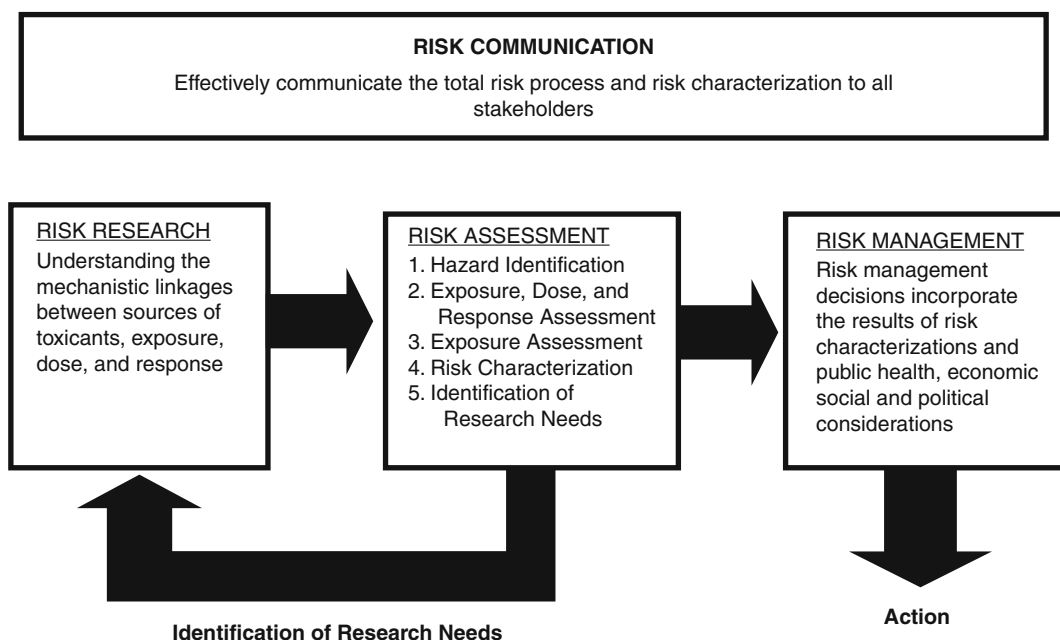
Quality and, perhaps, over time monitoring can establish whether the intended improvements in air quality are being achieved. The total air quality management cycle shown should be viewed as an integrated system. The system is dynamic with new sources being introduced, old sources retrofitted with controls, and some sources removed. In addition, as I will discuss later, the guidelines and standards are periodically reviewed and the concentration levels reduced. As some would say, the goal posts are constantly being moved to lower concentrations making it more challenging to achieve success and declare victory. As will be discussed later, how low is low enough?

Another key concept embedded within the air quality management system is the linkage from sources to health responses illustrated in Fig. 2. The schematic

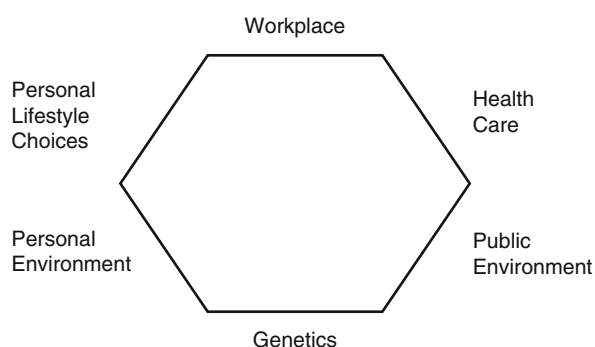
rendering in the figure is a general concept not exclusively linked to air pollution. The goal of air quality management is to have a positive impact on human health responses, the endpoint at the right of the figure. As illustrated in Fig. 3, multiple factors, not just air quality, influence human health. Thus, in a specific sense, the goal is to minimize the health impacts of air pollution relative to all the other factors that influence human health while achieving an overall improvement in health. Implicit in the setting of air quality standards and guidelines is the treatment of specific concentrations of agents in ambient air as surrogates for human health. Air quality can be readily characterized with an appropriately designed monitoring scheme for multiple air pollutants. In contrast, it is challenging to attempt to determine the association between specific risk factors, such as air pollution, and the health response of an individual. This is especially the case for current ambient concentrations of air pollutants in industrialized countries, concentrations substantially reduced from those observed only a few decades ago. Even when the focus of research investigations shifts from individuals to large populations, it is a challenge to characterize an association between a specific air pollutant and increased probability of a specific kind of response compared to a baseline. It is rarely possible to establish a causal relationship.

The complexity of relationships between sources of air pollution and health responses can be readily





**Air Quality Guidelines and Standards. Figure 2**  
The risk paradigm



**Air Quality Guidelines and Standards. Figure 3**  
Multiple factors influencing the health status of individuals and populations

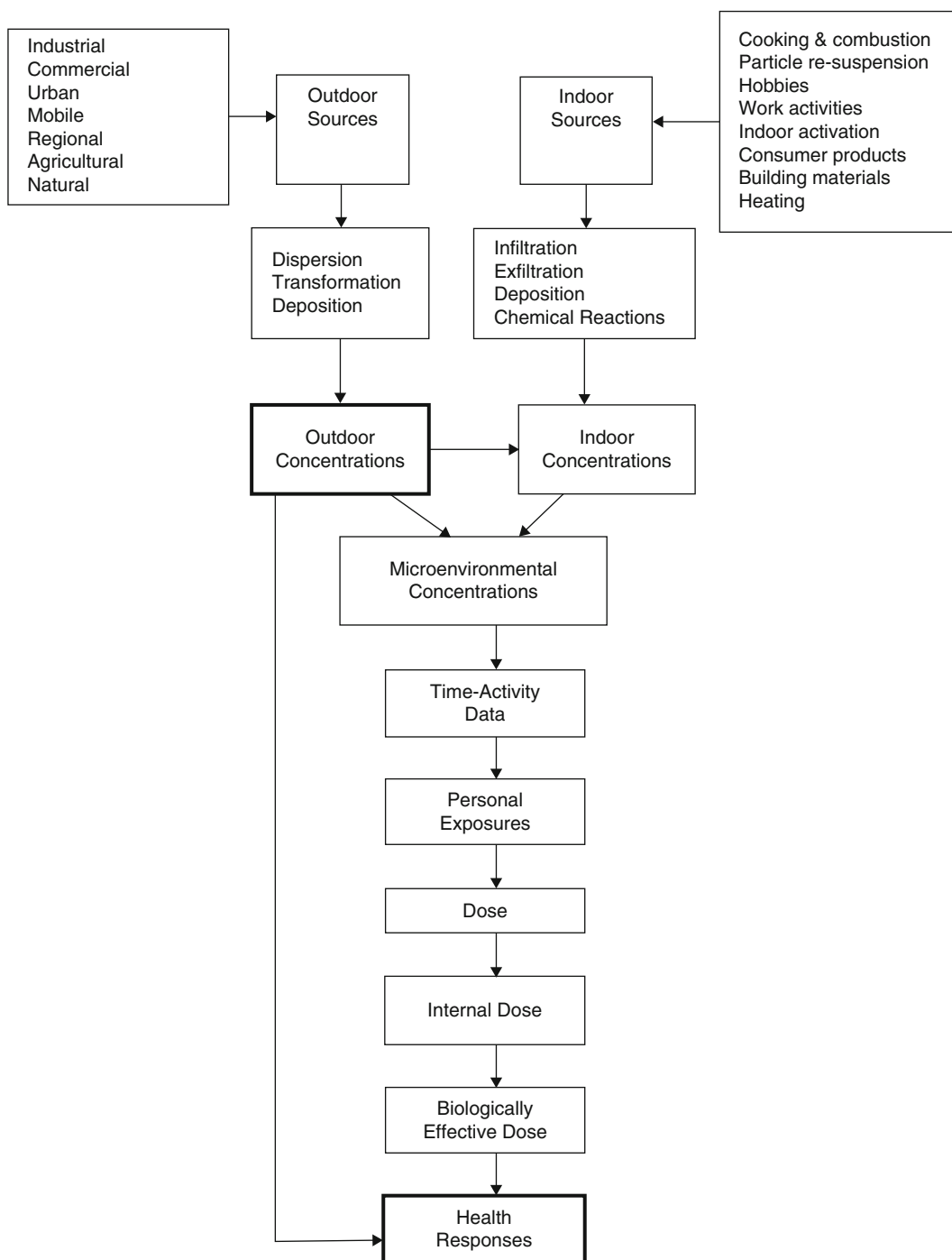
grasped by considering Fig. 4. In this figure, the concept of outdoor versus indoor exposures is introduced. The situation with regard to indoor exposures is even more complicated than may initially appear. This is the case because most individuals, during the course of their daily lives, spend variable amounts of time in multiple indoor environments: home (their bedroom, kitchen, living room, etc.), workplace, school, automobile, etc. Each of these micro-environments may have a different mix of air pollutants, both as to kind and

concentration, arising from multiple sources. Some variable portion of the indoor pollution may also arise from outdoors, the ambient environment.

The outdoor air pollution arises from multiple sources, some local, some regional, and some quite remote. Indeed, it is increasingly apparent that emissions arising in one continent such as Asia soon reach North America and then Europe. The air trajectory of emissions from coal-fired power plants in the Midwest extending to the Northeastern United States is well known. Dust originating from deserts in China regularly reaches the United States and Europe. Dust originating from the Sahara Desert also regularly circulates around the globe.

### Context for Air Quality Impacts

Although the focus of this entry is on air quality and its influence on human health, it is important that context be provided for considering the role of air pollution on human health. The introductory chapter by Ayres et al. [5] in the textbook on *Environmental Medicine* [6] provides a useful review of the multiple factors influencing health. As noted earlier (Fig. 3) multiple factors influence the health of each individual and, in



**Air Quality Guidelines and Standards. Figure 4**

Schematic rendering of the relationship between outdoor and indoor sources, personal exposure, biologically effective dose and health responses (Adapted from Özkaynak [50])

turn, the health of populations. Each individual has its own unique genetic background inherited from its parents. That genetic background can have a profound influence on the individual's potential health status throughout their life. Each of the other factors shown in the diagram can have a significant influence on the health status of an individual. Income and, more broadly, socioeconomic status are important factors influencing health.

The role of income and environmental factors are intertwined. This is readily apparent from considering Table 1 [31]. This table reports an analysis of the health impact of various risk factors in low- and middle-income countries versus high-income countries [31]. Urban air pollution and indoor smoke from household use of solid fuels are identified as risk factors. Taken together these two air pollution risk factors account for 5.2% of deaths in low- and middle-income countries versus 1.0% of deaths in high-income countries, the latter related exclusively to urban air pollution.

It is important to recognize that the role of socioeconomic status influencing health is a complex issue. Steenland et al. [59] have analyzed this issue for the US population and their findings are summarized in Table 2. Several points can be made using these findings. First, even for a country like the United States, typically identified as high-income, there are gradients in socioeconomic status and these have associated substantial gradients in health outcomes. Keep in mind the relative risks on the order of 2.0 when the relative risks of particulate matter and ozone are discussed. It is obvious that the national good will be served by improving the socioeconomic status of the total population. Second, a multitude of factors have a role in influencing the health status of the individuals in the lower quartile. Poor air quality to the extent it is experienced more by individuals in the lower quartile than in the upper quartile may be one of those factors. The substantial influence of socioeconomic status on health emphasizes the importance of attempting to control for it as a variable in epidemiological studies of air pollution and to report findings on it to provide perspective for the reported effects of air pollution.

Figure 5 provides a related perspective by contrasting environmental with non-environmental contributions to specific categories of disease using

disability-adjusted life years as an endpoint [55]. The merits of monetizing risks and benefits as statistical life-years versus lives have been reviewed by Sunstein [61]. Of special note are four diseases: lower respiratory infections, chronic obstructive pulmonary disease (COPD), ischemic heart disease, and lung cancer. The first three of these conditions account for about 4% of the total global burden disease in disability-adjusted life years while lung cancer accounts for about 1%. Air pollution is generally viewed as a contributor to global disease burden. However, its contribution needs to be placed in perspective relative to other major contributors to ill health.

In this entry, there is a focus on the development of air quality standards in the United States since, to a large degree, these activities have served as a template for other countries. For comparative purposes, the World Health Organization's Air Quality Guidelines have also been summarized. Reference will be made at several junctures to the distinction between guidelines providing guidance for policymakers versus standards that are legally binding and, in turn, may mandate certain actions. The omission of detailed discussions of guidelines and standards as developed and used in other countries around the globe is not intended as a judgment of their quality or utility.

### The US Legal Framework for Air Quality Standards

As concern mounted for the poor quality of air, citizens increasingly turned to the government to enact laws to restrict emissions and improve air quality [7, 60]. Bachmann's perspective is of special interest since as a US EPA employee for more than three decades, he had a central role in the setting of NAAQS. Readers interested in legal details of the Clean Air Act and NAAQS will find the summary of Martineau and Novello [33] useful. Initially, these air quality initiatives were focused at the local level in major cities around the world. In the United States, heavily industrialized cities, such as Pittsburgh, Chicago, Philadelphia, and Detroit, enacted legislation setting local standards. In most cases, these early standards focused on smoke.

By the mid-twentieth century in the United States, it had become apparent that local efforts on air

**Air Quality Guidelines and Standards. Table 1** Deaths by cause: low- and middle-income countries, high-income countries, and world, 2001

|                                                            | Low- and middle-income deaths | High-income deaths | World deaths  |
|------------------------------------------------------------|-------------------------------|--------------------|---------------|
| Total number (thousands)                                   | 48,351                        | 7,891              | 56,242        |
| Rate per 1,000 population                                  | 9.3                           | 8.5                | 9.1           |
| Age-standardized <sup>a</sup> rate per 1,000               | 11.4                          | 5.0                | 10.0          |
| Risk factor                                                | Number in thousands (percent) |                    |               |
| Childhood and maternal undernutrition                      |                               |                    |               |
| Childhood underweight                                      | 3,630 (7.5)                   | 0 (0.0)            | 3,630 (6.5)   |
| Iron-deficiency anemia                                     | 613 (1.3)                     | 8 (0.1)            | 621 (1.1)     |
| Vitamin A deficiency                                       | 800 (1.7)                     | 0 (0.0)            | 800 (1.4)     |
| Zinc deficiency                                            | 849 (1.8)                     | 0 (0.0)            | 849 (1.5)     |
| Other nutrition-related risk factors and Physical activity |                               |                    |               |
| High blood pressure                                        | 6,223 (12.9)                  | 1,392 (17.6)       | 7,615 (13.5)  |
| High cholesterol                                           | 3,038 (6.3)                   | 842 (10.7)         | 3,880 (6.9)   |
| Overweight and obesity                                     | 1,747 (3.6)                   | 614 (7.8)          | 2,361 (4.2)   |
| Low fruit and vegetable intake                             | 2,308 (4.8)                   | 333 (4.2)          | 2,641 (4.7)   |
| Physical inactivity                                        | 1,559 (3.2)                   | 376 (4.8)          | 1,935 (3.4)   |
| Addictive substances                                       |                               |                    |               |
| Smoking                                                    | 3,340 (6.9)                   | 1,462 (18.5)       | 4,802 (8.5)   |
| Alcohol use                                                | 1,869 (3.9)                   | 24 (0.3)           | 1,893 (3.4)   |
| Illicit drug use                                           | 189 (0.4)                     | 37 (0.5)           | 226 (0.4)     |
| Sexual and reproductive health                             |                               |                    |               |
| Unsafe sex                                                 | 2,819 (5.8)                   | 32 (0.4)           | 2,851 (5.1)   |
| Nonuse and use of ineffective methods of contraception     | 162 (0.3)                     | 0 (0.0)            | 62 (0.3)      |
| Environmental risks                                        |                               |                    |               |
| Unsafe water, sanitation, and hygiene                      | 1,563 (3.2)                   | 4 (<0.1)           | 1,567 (2.8)   |
| Urban air pollution                                        | 735 (1.5)                     | 76 (1.0)           | 811 (1.4)     |
| Indoor smoke from household use of solid fuels             | 1,791 (3.7)                   | 0 (0.0)            | 1,791 (3.2)   |
| Other selected risks                                       |                               |                    |               |
| Contaminated injections in health care setting             | 407 (0.8)                     | 4 (<0.1)           | 412 (0.7)     |
| Child sexual abuse                                         | 65 (0.1)                      | 6 (<0.1)           | 71 (0.1)      |
| All selected risk factors together                         | 22,014 (45.6)                 | 3,473 (44.0)       | 25,488 (45.3) |

<sup>a</sup>Age-standardized using the WHO World Standard Population

Source: Adapted from [31]

**Air Quality Guidelines and Standards. Table 2** The impact of socioeconomic status on mortality

| Mortality rate ratio = $\frac{\text{lowest quartile}}{\text{highest quartile}}$ of socioeconomic status |                               |                  |
|---------------------------------------------------------------------------------------------------------|-------------------------------|------------------|
| Mortality                                                                                               | Men                           | Women            |
| All cause                                                                                               | 2.02 (1.95–2.09) <sup>a</sup> | 1.29 (1.25–1.32) |
| Heart disease                                                                                           | 1.88 (1.83–1.93)              | 1.84 (1.76–1.93) |
| Stroke                                                                                                  | 2.25 (2.14–2.37)              | 1.53 (1.44–1.62) |
| Diabetes                                                                                                | 2.19 (2.07–2.32)              | 1.85 (1.72–2.00) |
| COPD                                                                                                    | 3.59 (3.35–3.83)              | 2.09 (1.91–2.30) |
| Lung cancer                                                                                             | 2.15 (2.07–2.23)              | 1.31 (1.25–1.39) |
| Breast cancer                                                                                           | –                             | 0.76 (0.73–0.79) |
| Colorectal cancer                                                                                       | 1.21 (1.16–1.27)              | 0.91 (0.86–0.96) |
| External causes                                                                                         | 2.67 (2.58–2.78)              | 1.41 (1.35–1.48) |

<sup>a</sup>95% confidence interval

Source: Steenland et al. [59]

pollution control were not sufficient, and complementary efforts were needed at the state or national level. At the national level in the United States post-WWII, the lead agency for air quality was the Division of Air Pollution of the Public Health Service within the Department of Health, Education and Welfare. The Air Pollution Control Act of 1955 [60] provided funding to the PHS for air pollution control research and to support the state's efforts. However, the impact of this Act on progress in managing air quality was limited.

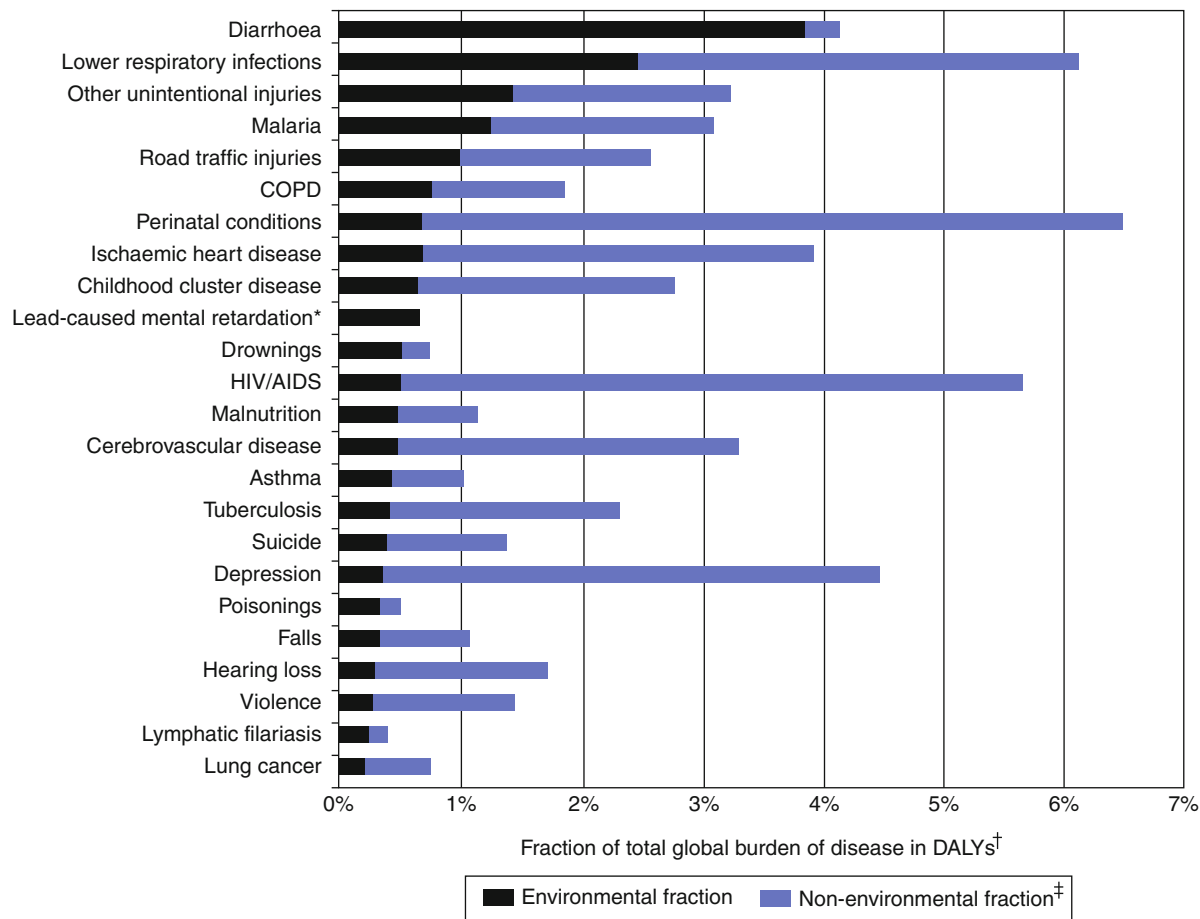
In 1963, the Clean Air Act (CAA) was passed, which became the principal national statute in the United States concerned with air quality [14]. The original CAA [14] directed the then Department of Health, Education and Welfare (HEW) to “compile and publish criteria on the effects of air pollutants,” hence the identification of “criteria pollutants” and “criteria documents” summarizing the scientific knowledge on certain air pollutants arising from multiple sources and found across the United States. The National Air Pollution Control Administration (NAPCA) within HEW was initially assigned responsibility of administering the CAA. The role of NAPCA was strengthened with the CAA Amendments of 1967. However, in the absence of a specific time schedule progress remained slow. A single document on SO<sub>2</sub> was developed [32]. When the US Environmental Protection Agency (EPA)

was created in 1970 by consolidating numerous existing federal agencies, responsibility for administering the CAA was effectively transferred from NAPCA to the EPA.

In 1970, almost coincident with creation of the US Environmental Protection Agency, major amendments to the CAA were passed that opened a new era in managing air quality in the United States [15]. The Clean Air Act as amended is a very broad piece of legislation. The sections dealing with the establishment of National Ambient Air Quality Standards are a key part of the Act and are viewed by many individuals as the core of the Act. Other sections of the Act provide authority for implementing processes to achieve attainment of the NAAQS and otherwise in improving air quality. This includes a delegation of responsibility to the individual States for taking steps to attain the NAAQS. This includes the development of State Implementation Plans if the States have areas that are not in attainment with regard to any specific NAAQS as evaluated by the concentration and statistical form of the Standard.

Other sections of the Act address both Stationary and Mobile Sources and the control of their emissions. This includes specific emission standards for vehicles and provision for issuance of fuel efficiency goals that in the course of being achieved impact on air quality. The rigid “command and control” and





COPD, Chronic obstructive pulmonary disease.  
\* Lead-caused mental retardation is defined in the World Health Organization list of diseases for 2002, accessed at: [www.who.int/evidence](http://www.who.int/evidence).  
<sup>†</sup> DALYs represent a weighted measure of death, illness and disability.  
<sup>‡</sup> For each disease, the fraction attributable to environmental risks is shown in black, Grey plus black represents the total burden of disease.

**Air Quality Guidelines and Standards. Figure 5**  
Diseases with the largest environmental contribution. *AIDS* acquired immune deficiency syndrome, *DALY* disability-adjusted life-year, *HIV* human immunodeficiency virus (Adapted from Pruss-Ustun and Corvalan, with permission)

standard-setting approach of the United States contrasts sharply with the guidance approach taken by the World Health Organization (WHO).

**National Ambient Air Quality Standards**

Section 108 of the amended CAA specifies the development of scientific criteria for certain air pollutants. It specifies that the Administrator publish a list for

pollutants: (a) which have an adverse effect on public health or welfare, and (b) which are derived from numerous or diverse mobile or stationary sources. It explicitly states – “the criteria shall accurately reflect the latest scientific knowledge on the kind and extent of all identifiable effects on public health and welfare which may be expected from the presence of the pollutant in ambient air, in varying quantities” [15].

Section 109 of the CAA [15] directs the Administrator to propose and promulgate “primary” and “secondary” NAAQSs for criteria pollutants identified under Section 108. The primary Standards are to be set to protect public health; secondary Standards are to be set to protect the public welfare such as effects on soils, water, crops, visibility, and deterioration of property. In this entry, I focus on the setting of the primary Standards and on the specific NAAQS. However, the issues discussed are also broadly applicable to the setting of secondary Standards.

Section 109(b)(1) defines a primary NAAQS as one that “the attainment and maintenance of which in the judgment of the Administrator, based on the criteria and allowing an adequate margin of safety, is requisite to protect the public health.” The margin of safety, as interpreted by the EPA, is intended to address uncertainties associated with inconclusive scientific and technical information at the time the Standard is set and to account for hazards that research has not yet identified.

The primary Standards are intended to protect against “adverse effects, not necessarily against all identifiable effects of changes produced by a pollutant.” Although the US Congress did not rigorously define an adverse effect, it did provide general guidance in the legislative history of the debate on the CAA [30]. Congress was concerned with effects ranging from cancer, metabolic and respiratory disease, and impairment of mental processes to headaches, dizziness, and nausea.

Congress also noted concern for sensitive population groups in setting the NAAQSs. In particular, Congress noted that the Standards should protect “particularly sensitive citizens such as bronchial asthmatics and those with emphysema who in the normal course of daily activity are exposed to the ambient environment.” This has been interpreted to exclude individuals who are not performing normal activities, such as individuals who are hospitalized. Further guidance was given noting that the Standard is statutorily sufficient whenever there is “an absence of adverse effect on the health of a statistically related sample of persons in sensitive groups from exposure to the ambient air.”

The challenge of interpreting the “adequate margin of safety” language of the CAA was noted in an editorial by Donald Kennedy on *Risk versus Risk* published when

he served as Editor-in-Chief of Science [27]. He wrote – “In the United States and some other industrial democracies, where people and their governments tend to be *risk averse*, legislatures, courts, and administrative entities usually create a presumption favoring more safety rather than less. The definitions of risk in law are often vague (‘reasonable certainty of no harm’ or ‘adequate margin of safety’) and are likely to encourage an *unrealistic belief* that risks can be minimized or even eliminated altogether.” I think Kennedy has captured the conundrum posed by the language of the CAA, a conundrum that has been addressed by US Supreme Court Justice Stephen Breyer as I will relate later.

In 1977, several key amendments were made to the CAA [16]. Concern about slow action of the EPA in preparing criteria documents and reassessing NAAQSs prompted a legislated requirement that the NAAQSs be reevaluated not later than January 1, 1980, and at 5-year intervals thereafter. Reevaluation was not intended to automatically result in changes in the NAAQSs for a pollutant; rather, reevaluation was intended to ensure that the scientific database was reviewed and that the NAAQSs were consistent with current knowledge. To my knowledge, this requirement for mandatory review every 5 years is unique to the setting of the NAAQS in the United States. Indeed, I know of no other statute calling for an updating of the science and reconsideration of the Standard every 5 years.

Peer review of the earliest criteria documents prepared by the EPA was carried out by various committees of the agency’s Science Advisory Board. The 1977 amendments to the CAA institutionalized the peer-review process for the NAAQS [16]. The amendment requires the EPA Administrator to appoint an independent scientific committee, composed of seven members, including at least one member of the National Academy of Sciences, one physician, and one person representing state air pollution control agencies to advise the Administrator on the science informing the policy judgments made in setting the NAAQS. The EPA has implemented this provision of the CAA by appointing a Committee, which designated itself as the Clean Air Scientific Advisory Committee (CASAC). The CASAC is directly responsible to the EPA Administrator, although it functions administratively as one of the standing committees of the

EPA Science Advisory Board. Traditionally, the requirement for one CASAC member to be a member of the National Academy of Sciences has been broadly interpreted to also include membership in either the National Academy of Engineering or the Institute of Medicine. To complement the expertise of regular members of the CASAC, a number of consultants, sometimes as many as a dozen, with specialized expertise usually have been added to the review panels for specific pollutants.

The CAA was amended again in 1990 [17]. Although major changes were made in the CAA with these amendments, especially with regard to the regulation of hazardous air pollutants, there were no changes in the fundamental approach to dealing with the setting of NAAQS for criteria pollutants. However, there were changes in the CAA that have had major impact on the regulation of emissions of PM and precursors especially from large power plants.

The pollutants originally designated as “criteria pollutants” because of their ubiquitous distribution and potential to endanger health were photochemical oxidants (later regulated as ozone), particulate matter (later regulated as Total Suspended Particulates, then

PM<sub>10</sub>, and PM<sub>2.5</sub>), carbon monoxide, sulfur oxides (regulated as sulfur dioxide), nitrogen oxides (regulated as NO<sub>2</sub>), and non-methane hydrocarbons (later dropped as a criteria pollutant). The EPA [64] established NAAQS for these pollutants soon after the Agency was created using existing scientific documentation, that is, criteria. The original NAAQS set in 1971 are shown in Table 3. As I will discuss below, the EPA later added Lead as a criteria pollutant with legal prodding from the National Resources Defense Council [49, 72].

### US Process for Setting National Ambient Air Quality Standards

The process for developing and issuing NAAQS after the initial NAAQS were set is quite complex as discussed by Bachmann [7] and McClellan [36, 37, 39, 41]. Key elements of the process, as used until quite recently, include preparation and review of (a) criteria document, (b) staff paper, (c) more recently a risk assessment, and (d) a regulatory decision package leading to the Administrator’s policy judgment decisions as to the proposed and final NAAQS which are

**Air Quality Guidelines and Standards. Table 3** Primary and secondary National Ambient Air Quality Standards initially set by the US Environmental Protection Agency in 1971<sup>a</sup>

| Pollutant (indicator) <sup>b</sup>          | Level                             | Averaging time | Form <sup>b</sup>                          |
|---------------------------------------------|-----------------------------------|----------------|--------------------------------------------|
| PM (total suspended particulates)           | 75 µg/m <sup>3</sup>              | Annual         | Geometric mean                             |
|                                             | 260 µg/m <sup>3</sup>             | 24 h           | Not to be exceeded more than once per year |
| TSP secondary standard                      | 150 µg/m <sup>3</sup>             | 24 h           | Not to be exceeded more than once per year |
| SO <sub>2</sub>                             | 0.03 ppm                          | Annual         | Arithmetic mean                            |
|                                             | 0.14 ppm                          | 24 h           | Not to be exceeded more than once per year |
| SO <sub>2</sub> secondary standards         | 60 µg/m <sup>3</sup> (0.02 ppm)   | Annual         | Arithmetic mean                            |
|                                             | 1,300 µg/m <sup>3</sup> (0.5 ppm) | 3 h            | Not to be exceeded more than once per year |
| CO                                          | 10 µg/m <sup>3</sup> (9 ppm)      | 8 h            | Not to be exceeded more than once per year |
|                                             | 40 µg/m <sup>3</sup> (35 ppm)     | 1 h            | Not to be exceeded more than once per year |
| Photochemical oxidants (as O <sub>3</sub> ) | 200 µg/m <sup>3</sup> (0.08 ppm)  | 1 h            | Not to be exceeded more than once per year |
| HC                                          | 160 µg/m <sup>3</sup> (0.24 ppm)  | 3 h, 6–9 a.m.  | Not to be exceeded more than once per year |
| NO <sub>2</sub>                             | 100 µg/m <sup>3</sup> (0.053 ppm) | Annual         | Arithmetic mean                            |

<sup>a</sup>With the exception of PM and SO<sub>2</sub>, secondary NAAQS set identical to the primary standards

<sup>b</sup>The terms “indicator” and “form” were not used in the 1971 notice

published in the Federal Register. Traditionally, CASAC focused its attention on reviewing the Criteria Documents and Staff Papers and, more recently, a formal Risk Assessment. As an aside, the process was changed at the end of 2006 [51] with an Integrated Science Assessment and Policy Assessment Document replacing the Criteria Document and Staff Paper. Time will tell if these changes really improve the overall process.

In addition to the documents noted above, the Agency now prepares a Regulatory Impact Analysis which is required under Executive Order 12866 issued by President Clinton [18] that applies to economically significant rules that have “an annual effect on the economy of \$100 million or more or adversely effect in a material way the economy, a sector of the economy, productivity, competition, jobs, the environment, public health or safety or site, local, or tribal governments or communities.” The results of the Regulatory Impact Analysis are not considered during the NAAQS rule-making process given the prohibition of consideration of cost in the setting of the NAAQS [93].

It is important to recognize that the NAAQS process in the United States started as an informal rule-making process [7, 41, 52]. The DC Court of Appeals found that the informal process was not sufficiently well documented with regard to the four key elements (indicator, averaging time, concentration, and statistical form) that the Court could perform a judicial review of the rules that were promulgated. This resulted in a judicial decision to overturn the secondary SO<sub>2</sub> NAAQS in 1973 [10]. This in turn led the US Environmental Protection Agency to develop more rigorous procedures including formal documentation [52].

Bachmann [7] summarized the adoption of these procedures in the late 1970s noting the following points: “(1) EPA was to make available to the public the information and technical methodologies it relied upon by the time of proposal; (2) the preambles to proposal and final rules were to provide a detailed explanation of EPA’s decision; (3) EPA was required to respond to all “significant” comments on the proposal by the time it issues its final rule; and (4) all of the above documents, analyses, preambles, and responses constituted the record that the court would examine in reviewing the final standard decision. Objections not raised in the record could not be raised in court. The

halcyon days of a speedy NAAQS process were over.” I agree that the speed of the process was reduced, however, I would add that transparency of the process was also substantially improved. Congress apparently agreed and these provisions were substantially codified by the CAA Amendments of 1977 [16].

EPA’s implementation of the CAA, especially its setting of NAAQS even with improved documentation, has been a matter of continuing controversy and litigation. Some persons might note that controversy and litigation were enhanced by improved documentation. Bachmann [7] summarizes many of the key legal cases in his review. In this entry, I will only highlight several key legal cases. The 1997 revisions of the Ozone NAAQS [65] and Particulate Matter NAAQS [66] proved to be very contentious including within CASAC. For example, the members of the CASAC PM Panel, chaired by George Wolff, had a range of views on the PM<sub>2.5</sub> standard that was being set for the first time replacing, in part, the PM<sub>10</sub> standard. This range of views was clearly articulated in the CASAC Chair’s letter [97] to the Administrator by including a Table showing the views of each individual member of the CASAC Panel. This approach stands in stark contrast to the approach of more recent CASAC Panels that have emphasized consensus advice.

The contentious nature of the debate over these revised NAAQS prompted Administrator Browner to involve President Clinton. Bachmann [7] recounts that Administrator Browner had a 1-h meeting on these standards with the President – “she reported that the President quickly accepted her decision and spent much of the time discussing how to reduce unnecessary burdens in the implementation process. This resulted in some of us writing the first draft of a letter that was later sent by President Clinton [19] to EPA directing implementation be carried out so as to “maximize common sense, flexibility, and cost effectiveness.” Not surprisingly, President Clinton [46] had a role in announcing the tighter standards which included for the first time a separate PM<sub>2.5</sub> standard to complement the PM<sub>10</sub> standard and a shift from a 1-h averaging time to an 8-h averaging time standard for Ozone.

The issuance of a revised PM NAAQS triggered the legal case of *American Trucking Associations v. EPA* [4]. The Court found “the growing empirical evidence demonstrating a relationship between fine particle

pollution and adverse health effects amply justifies establishment of new fine particulate standards.” The Court went on to find “ample support” for EPA’s decision to regulate coarse particulate pollution, but vacated the 1997 PM<sub>10</sub> standards, concluding in part that PM<sub>10</sub> is a “poorly matched indicator for coarse particulate pollution” because it includes fine particulates. Subsequently, EPA removed the vacated 1997 PM<sub>10</sub> standard allowing the 1987 PM<sub>10</sub> standard to remain in place along with the new PM<sub>2.5</sub> issued in 1997.

In addition, the three-judge panel held, two to one, that EPA’s approach to setting the level of the PM and Ozone standards in 1997 effected “an unconstitutional delegation of legislative authority.” The Panel found that “the factors EPA uses in determining the degree of public health concern associated with different levels of ozone and particulate matter are reasonable.” However, it remanded the rule to EPA. The Judicial Panel stated that when the Agency considers these factors for potential non-threshold pollutants “what EPA lacks is any determinate criterion for drawing lines” to determine the level at which the standards should be set. The Panel also found that the Administrator, under the CAA, is not permitted to consider the cost of implementing these standards in setting them.

Not surprisingly, the nature of the Circuit Court opinion resulted in cross appeals being filed on the several issues and the involvement of the US Supreme Court. The Supreme Court in February 2001 issued a unanimous opinion upholding EPA’s position on both the constitutional and cost issues [93]. On the constitutional issue, the Supreme Court held that the statutory requirement that the NAAQS be “requisite” to protect public health with an adequate margin of safety sufficiently guided EPA’s discretion, affirming EPA’s approach of setting standards that are neither more nor less stringent than necessary.

Supreme Court Justice Breyer, who participated in the *Whitman v. American Trucking Associations* case, is well known and highly regarded for his opinions and writings on risk assessment [11, 12]. Thus, it is not surprising that he took the opportunity in *Whitman v. American Trucking Associations* [93] to offer comments on the standard-setting process and, specifically, the identification of the level of the NAAQS and the associated level of health risk. While concurring that EPA cannot consider the costs of implementing the

Standards, he went on to note – “this interpretation of §109 does not require the EPA to eliminate every health risk, however slight, at any economic cost, however great, to the point of “hurtling” industry over “the brink of ruin,” or even forcing “deindustrialization.” (Id. At 494) (Breyer, J., concurring in part and concurring in judgment) (citations omitted). Rather, as Justice Breyer explained:

- The statute, by its express terms, does not compel the elimination of all risk; and it grants the Administrator sufficient flexibility to avoid setting ambient air quality standards ruinous to industry.

Section 109(b)(1) directs the Administrator to set standards that are “requisite to protect the public health” with “an adequate margin of safety.” But these words do not describe a world that is free of all risk – an impossible and undesirable objective (citation omitted). Nor are the words “requisite” and “public health” to be understood independent of context. We consider football equipment “safe” even if its use entails a level of risk that would make drinking water “unsafe” for consumption. And what counts as “requisite” to protecting the public health will similarly vary with background circumstances, such as the public’s ordinary tolerance of the particular health risk in the particular context at issue. The Administrator can consider such background circumstances when “deciding what risks are acceptable in the world in which we live.” (Citation omitted).

The statute also permits the Administrator to take account of comparative health risks. That is to say, she may consider whether a proposed rule promotes safety overall. A rule likely to cause more harm to health than it prevents is not a rule that is “requisite to protect the public health.” For example, as the Court of Appeals held and the parties do not contest, the Administrator has the authority to determine to what extent possible health risks stemming from reductions in tropospheric ozone (which, it is claimed, helps prevent cataracts and skin cancer) should be taken into account in setting the ambient air quality standard for ozone. (Citation omitted).

The statute ultimately specifies that the standard set must be “requisite to protect the public health” “in the judgment of the Administrator,” §109(b)(1), 84 Stat. 1680 (emphasis added), a phrase that grants the Administrator considerable discretionary standard-setting authority.



The statute's words, then, authorize the Administrator to consider the severity of a pollutant's potential adverse health effects, the number of those likely to be affected, the distribution of the adverse effects, and the uncertainties surrounding each estimate (citation omitted). They permit the Administrator to take account of comparative health consequences. They allow her to take account of context when determining the acceptability of small risks to health. And they give her considerable discretion when she does so.

The discretion would seem sufficient to avoid the extreme results that some of the industry parties fear. After all, the EPA, in setting standards that "protect the public health" with "an adequate margin of safety," retains discretionary authority to avoid regulating risks that it reasonably concludes are trivial in context. Nor need regulation lead to deindustrialization. Pre-industrial society was not a very healthy society; hence a standard demanding the return of the Stone Age would not prove "requisite to protect the public health."

Although I rely more heavily than does the Court upon legislative history and alternative sources of statutory flexibility. I reach the same ultimate conclusion, Section 109 does not delegate to the EPA authority to base the National Ambient Air Quality Standards, in whole or in part, upon the economic costs of compliance.

The case of *Whitman v. American Trucking Associations* [93] is widely cited for the conclusion that EPA cannot consider the economic costs of compliance in the setting of NAAQS. Unfortunately, in my opinion, insufficient attention is given to the thoughtful guidance of Justice Breyer on exercising policy judgment informed by science in deciding on an acceptable level of health risk, a judgment that in turn determines the level and statistical form of each NAAQS. It is interesting that Justice Breyer's opinion appeared in Administrator Johnson's notice of the Ozone NAAQS [81], but did not appear in Administrator Jackson's "reconsideration" proposal for Ozone [85] that will be discussed later.

## Paradigm Shift

At this juncture, it is appropriate to note that it is my view that a paradigm shift has taken place in the selection of the level of each NAAQS over the past four

decades. In my opinion, the paradigm shift has been driven in part by the changing nature of the scientific evidence of pollution effects. In the 1970s, most scientists and regulations viewed the criteria pollutants as having a threshold in their ambient concentration-response relationship, sometimes referred to as a hockey stick function [36, 39]. In contrast cancer-causing agents have been assumed to have linear, non-threshold, concentration-response relationships.

In the late 1960s and early 1970s, the available data on each criteria pollutant were quite modest, with attention focusing on only a few epidemiological studies. For those few studies, attention often focused on whether a relative risk on the order of 2.0 was statistically significant or not. For a given pollutant there were few, if any, controlled human exposure studies. The data from laboratory animal studies with specific pollutants had frequently been acquired in studies with exposure concentrations much higher than ambient concentrations of the pollutants. The results of these studies raised questions about extrapolation both from laboratory animals to humans and high to low exposure concentrations. The general approach taken was to identify the lowest levels where effects were statistically significant and assume this was the inflection point in the concentration-response relationship. It could be argued that setting the Standard at a lower concentration satisfied the requirement for "an adequate margin of safety."

In contrast, the most recent reviews have involved thousands of papers with observations ranging from human population level to studies of intact laboratory animals to studies of effects of air pollutants on cells and molecules. Despite the huge number of published studies, attention has ultimately focused on the results of a few studies where attention turns to the critical issue of how they inform the setting of the level and statistical form of the standard. In the epidemiological studies, the debate often focuses on whether relative risks of less than 1.1 for excess morbidity and mortality are significant. Of course, the specific relative number is dependent on the denominator being used, for example, per 10 ppb of ozone (24-h average). For controlled exposure clinical studies, attention has focused on the lowest levels with statistically significant changes, which in turn, triggers debate over whether the changes are adverse.

A news report [63] in *Science* that I view as a classic report highlighted the issues involved in the search for subtle links between diet, lifestyle, or environmental factors and disease, especially using retrospective observational studies. I especially liked the quote at the end attributed to Sander Greenland, a well-known epidemiologist, in offering advice to his “most sensible, level-headed, estimatable colleagues.” Remember, he says – “there is nothing sinful about going out and getting evidence, like asking people how much do you drink and checking breast cancer records. There’s nothing sinful about seeing if that evidence correlates. There’s nothing sinful about checking for confounding variables. The sin comes in believing a casual hypothesis is true because your study came up with a positive result, or believing the opposite because your study was negative.”

It is interesting to note that CASAC discussions of criteria pollutant effects have frequently focused initially on the level of the standard sometimes devoid of any consideration of the statistical form of the level. This approach was in keeping with traditional practice in the setting of standards such as Threshold Limit Values for occupational exposures to chemicals based on scientific judgments [34–36, 40]. That approach has involved a review of the available human data on a toxic chemical to determine a no-observed effect level, or the lowest observed effect level, and then using safety factors to arrive at an acceptable exposure level set at a lower level. In the absence of adequate human data, laboratory animal data are used and an additional safety factor used to account for the potential that the animal observations might not adequately predict human effects. This approach was routinely used for a wide range of health responses that were assumed to have an exposure-response relationship that exhibited either a true or practical threshold, an excess of effects above some level and an absence of effects below that level. A review of the earliest Criteria Documents and, indeed, also the Staff Papers, documents that a similar line of reasoning was used in the setting of the NAAQS.

The implementation of standards set with this approach soon revealed that if the standard was to be rigorously enforced, that is, no exceedances of the specific level of the standard, the practical effect would be to

cause average levels of the pollutant to be reduced to levels far below the standard so as to avoid the occasional high value exceeding the standard. Fortunately, common sense prevailed and the EPA, over time, moved to the practice of routinely linking the specific level of the standard with a statistical form such as the 98th percentile 24-h concentration averaged over 3 years or the fourth highest 8-h average concentration during a 3-year period. In my opinion, most of the attention of the CASAC in the NAAQS setting process has focused on the level of the standard with limited discussion of the statistical form of the Standard. In doing so, there has been a failure to recognize that the stringency of the standard and the degree of health protection provided depends on *both* the level and statistical form of the standard for a particular indicator and averaging time.

The challenges of selecting appropriate averaging times and statistical forms for the NAAQS are substantial. The original epidemiological and toxicological studies that provide the scientific information that should inform the setting of the NAAQS do not always report results with an averaging time that is the same as that ultimately selected for the Standard. Hence, the need to make extrapolations from results reported based on one metric, such as average daily exposure to a second metric such as an 8-h averaging time. The setting of standards at extreme values, the 98th percentile for NO<sub>2</sub> [86] or, even the 99th percentile form as done with the 1-h averaging time standard for SO<sub>2</sub> [87], results in extremely stringent standards that at best are only very loosely related to the underlying data. In my view, selection of the specific level and statistical form for a specific averaging time is clearly policy judgments beyond the dominion of the science. It is my opinion that decisions on the selection of specific averaging times are ultimately policy judgments reserved to the Administrator by the language of the CAA.

In the 1990s, concurrent with the increasingly widespread use of formal risk analysis procedures across society [35, 40], EPA moved to quantify the health benefits associated with setting the NAAQS at various levels, with an associated statistical form. I must admit to being an early advocate of formal quantification of health benefits. I viewed the approach then and I still do today, as providing useful guidance for making policy decisions. I did not envision that some enthusiastic

advocates of quantitative risk analysis would actually view the results of the analyses as being highly accurate projections of potential health benefits expressed to two or more significant figures, sometimes without any indication of uncertainty.

The quantification of health effects potentially associated with various levels and forms of the Standards requires several kinds of input. Examples are the risk assessments conducted by the USEPA for particulate matter [69], ozone [77, 78, 79] and sulfur dioxide [83]. Guidance has also been provided by the US National Research Council [48]. First and foremost, conduct of a risk assessment requires some knowledge of the nature of the ambient concentration-response relationships for various temporal metrics for the pollutant in question. Typically, the response term is expressed as excess risk per unit of increased concentration over some range of ambient concentrations. The question then becomes one of whether the relationship is linear and whether there is a threshold level below which the coefficient for excess risk holds or does not hold. The issue of whether there are or are not thresholds for non-cancer health endpoints from exposure to many agents is very contentious [56, 92]. It is well recognized that the use of linear no-threshold exposure-response relationships for cancer-causing agents is closely linked to experience with radiation [47]. An additional issue becomes the selection of suitable reference baseline statistics for the particular health effects to use along with a given ambient concentration-response coefficient. Yet another question becomes the appropriate population to be evaluated, a single city, multiple cities, or the population of the United States. It is obvious that there are substantial uncertainties associated with each component of the analysis of benefits of potential reductions in a given pollutant.

With the use of linear, no-threshold, ambient concentration-response models, the EPA has, on some occasions, calculated estimated excess morbidity and mortality effects attributable to the specific pollutant down to background concentrations. The Health Risk Assessment [77] and the Regulatory Impact Analysis [82] for the 2008 Ozone NAAQS [81] serve as examples. Further, dependent on the assumptions made with regard to how ambient concentrations of the pollutant would change in response to various levels

and forms of the standard, substantial health effects avoided, that is, health benefits can be calculated. Expressing the benefits in monetary units raises some critical issues. Typically, mortality effects drive the monetized benefits calculations because of the substantial value placed on deaths avoided. Sunstein [61] has addressed this contentious issue and has argued that years of life loss is a more appropriate metric than is the use of lives lost as is the typical approach used by the USEPA.

A key consideration as to whether these benefits can be realized relates to whether the roll-back in air concentrations that is assumed in the analysis as a result of the new standard can actually be realized. In part, this relates to how realistic the assumptions have been with regard to background levels. The analysis of benefits of reducing a given pollutant, say Ozone, becomes even more convoluted when the analysis includes postulated reductions in a second pollutant, such as particulate matter, associated with achieving the Ozone Standard. A discussion of this issue for Ozone can be found in McClellan et al. [43]. Indeed, as the levels of the standards are ratcheted down toward background levels, there is increasing uncertainty as to whether there are any health effects attributable to single pollutants and even greater uncertainty as to the magnitude of the health benefits associated with any new lower standard. The use of single pollutant models for estimating benefits raises the issue of double-counting of benefits as the benefits of the individual pollutants are aggregated.

Hence, the paradigm shift. It is apparent that in setting the earliest NAAQS some individuals, including CASAC members, envisioned that the standards were being set at levels protective of public health with an adequate margin of safety based on threshold ambient concentration-response models. In short, if there were health effects at the level and form of the NAAQS, they were *de minimis*. It is useful to recall the Editorial by Kennedy [27]. In contrast, more recent NAAQS have been set at levels which both the CASAC and EPA characterize as having residual health effects even if the standard were to be attained. The central question remains – how low is low enough? I view the answer as a policy judgment informed by science that can only be made by the EPA Administrator [41].

## WHO Process for Setting Air Quality Guidelines

The approach taken by the World Health Organization (WHO) contrasts with the USEPA approach. WHO was established in 1948 as a specialized Agency of the United Nations to be the directing and coordinating authority for international health matters and public health. WHO develops air quality guidelines (AQG) designed to offer guidance in reducing the health impacts of air pollution. The first AQG were produced in 1987 [94]. Updates were published in 2000 and 2005 [95, 96]. The AQG are intended to inform policymakers and to provide appropriate targets for a broad range of policy options for air quality management in different parts of the world, despite the more recent documents being titled AQG for Europe.

The WHO AQGs are developed by working groups of scientific experts [95, 96]. In addition to preparing AQG for the classical air pollutants (Particulate matter, Ozone, Nitrogen dioxide, and Sulfur dioxide) the WHO also developed AQG for (a) inorganic air pollutants, (b) certain indoor air pollutants, (c) polychlorinated biphenyls, dioxins, and furans, and (d) volatile organic compounds. The WHO AQG process does not include any opportunity for the involvement of interested parties such as the public, national governments, nongovernmental organizations, or industry. They represent expert judgment exercised by scientists knowledgeable of the subject material. This approach contrasts sharply with the more formal and complex approach followed by the US Environmental Protection Agency in developing NAAQS.

The primary aim [95] of the AQG is “to provide a uniform basis for the protection of public health and ecosystems from adverse effects of air pollution and to eliminate or reduce to a minimum exposure to those pollutants that are known or likely to be hazardous.” “They have the character of recommendations, and it is not intended or recommended that they simply be adopted as standards.” The WHO [95] report notes:

- In the process of moving from a “guideline” or a “guideline value” to a “standard,” a number of factors beyond the exposure-response relationship need to be taken into account. These factors include current concentrations of pollutants and exposure levels of a population, the specific mixture of air pollutants,

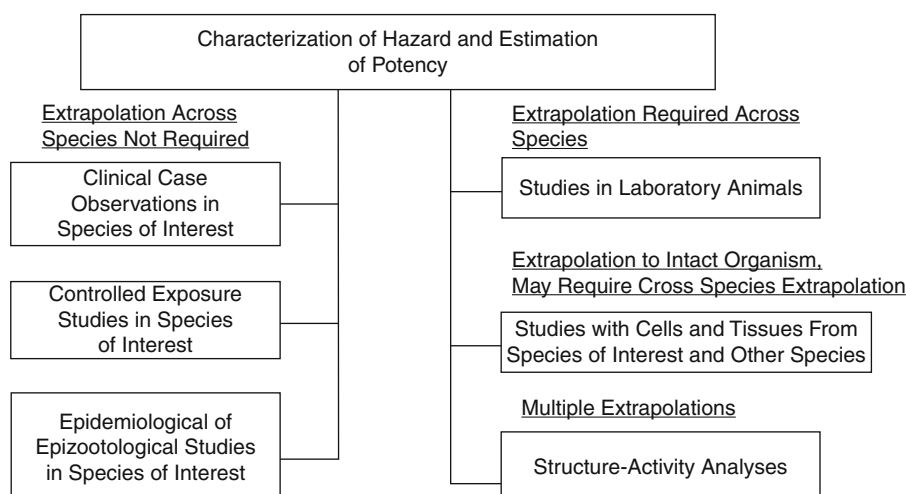
and the specific social, economic and cultural conditions encountered. In addition, the standard-setting procedure may be influenced by the likelihood of implementing the standard. These considerations may lead to a standard above or below the respective guideline value.

## Assessing Hazards and Risks

The risk approach used around the world today (Fig. 2) developed within the USEPA in parallel with the process for setting NAAQS [35, 36, 38, 39]. The setting of air quality standards and guidelines may be viewed as a specialized use of the risk characterization paradigm illustrated in Fig. 2 [35]. In particular, the establishment of air quality standards and guidelines draws on the schematic rendering in the box labeled “risk assessment.” An initial step in that process is to determine the hazard potential of the agent, in short, to determinewhether the specific agent has the potential for causing adverse health effects in humans. In the case of the criteria or common air pollutants discussed in this entry, there is sufficient evidence to establish that with exposure of sufficient duration and concentration the agent is capable of causing disease in people, based on epidemiological evidence. The key issue then becomes the nature of the relationship between ambient concentrations and duration of exposure and increased disease. The specific characteristics of that complex relationship are crucial to determining for a particular ambient concentration the estimated health impacts including the ambient concentration level. Moreover, the same information is required if one is to estimate the health benefits associated with a postulated improvement in air quality, that is, reductions in ambient concentrations.

## Sources of Information for Setting Standards and Guidelines

The multiple sources of information for evaluating the potential health effects of any air pollutant are illustrated in Fig. 6 [35–39, 41]. In the case of all the criteria air pollutants, the dominant sources of information are those based on the study of humans, the left side arm of the figure. However, all of the common air pollutants have been viewed as being sufficiently innocuous, and



**Air Quality Guidelines and Standards. Figure 6**  
Sources of information for evaluating potential health effects

with effects that are readily reversible when exposures are discontinued, that controlled human exposure studies can be conducted with the pollutant. Such studies have provided very useful information on short-term exposures, typically a few hours in duration as will be discussed later for ozone. A limitation of such studies is the appropriate requirement for the conduct of such studies only in healthy adults or individuals with modest preexisting disease. Controlled exposure studies in individuals with serious preexisting disease would not be ethical. The paramount advantage of controlled human exposure studies is that the exposures can be well characterized unlike the situation for field studies or epidemiological studies.

For some of the common air pollutants, valuable data on shorter-term exposures have been acquired from field studies of individuals exposed to ambient concentrations of pollutants. In these studies, the investigators can obviously only study air concentrations that are present and the extent they can be characterized and then relate the ambient concentrations to one or more health indices such as changes in lung function.

The predominant source of information on the health effects of the common air pollutants has been from epidemiological studies. It is beyond the scope of this entry to discuss the conduct and results of those studies in detail. The reader interested in more detail

will find the textbook on Environmental Medicine edited by Ayres et al. [6] useful. Suffice it to note that most epidemiological studies of air pollution are a kind of cohort study or a time-series study. None of the studies are strict cohort studies in which a population is randomly assigned to one or more treatment groups. This is rarely possible in studying air pollution. Investigators are constrained to studying populations under natural conditions of ambient exposure and other risk factors associated with the natural course of life. The study that may come closest to the ideal cohort study is the Six-City Study [20, 23] which will be discussed later. In such a study, the investigators may use regulatory compliance monitoring data supplemented by research monitoring to characterize ambient air concentrations of the various pollutants. Health outcomes information may be obtained by contact with the enrollees supplemented by information acquired from Administrative databases such as records of deaths. Other studies such as those using the ACS populations, for example, Pope et al. [54] are totally dependent upon regulatory compliance modeling and administrative databases for health outcomes. As might be expected, the quality of these studies is dependent upon the quantity and quality of the monitoring data, the availability of data on other risk factors, the duration of follow-up, the quality of the health outcome data, and the rigor of the analyses.



A second type of study design is that used in the time-series studies as pioneered by the Johns Hopkins University investigators in the National Morbidity and Mortality Air Pollution Studies (NMMAPS). These studies utilize regulatory compliance monitoring data to characterize ambient concentrations of pollutants and administrative databases for morbidity and mortality. These studies seek to characterize the relationship between ambient concentrations on a given day or series of days and excess morbidity or mortality occurring concurrently or on subsequent days. An example is the study of Bell et al. [8, 9] and Smith et al. [58] that will be discussed later. An advantage is that it is assumed that many other risk factors that may be influencing the health of the population is not varying on a day-to-day basis, for example, socioeconomic status. Other factors such as temperature may vary along with pollution levels.

In the next section, information on the health effects of particulate matter and ozone will be briefly reviewed as examples of the kind of information used in setting NAAQS. In this section I draw heavily on a previous review of risk assessment/risk management [40].

### Particulate Matter as an Example

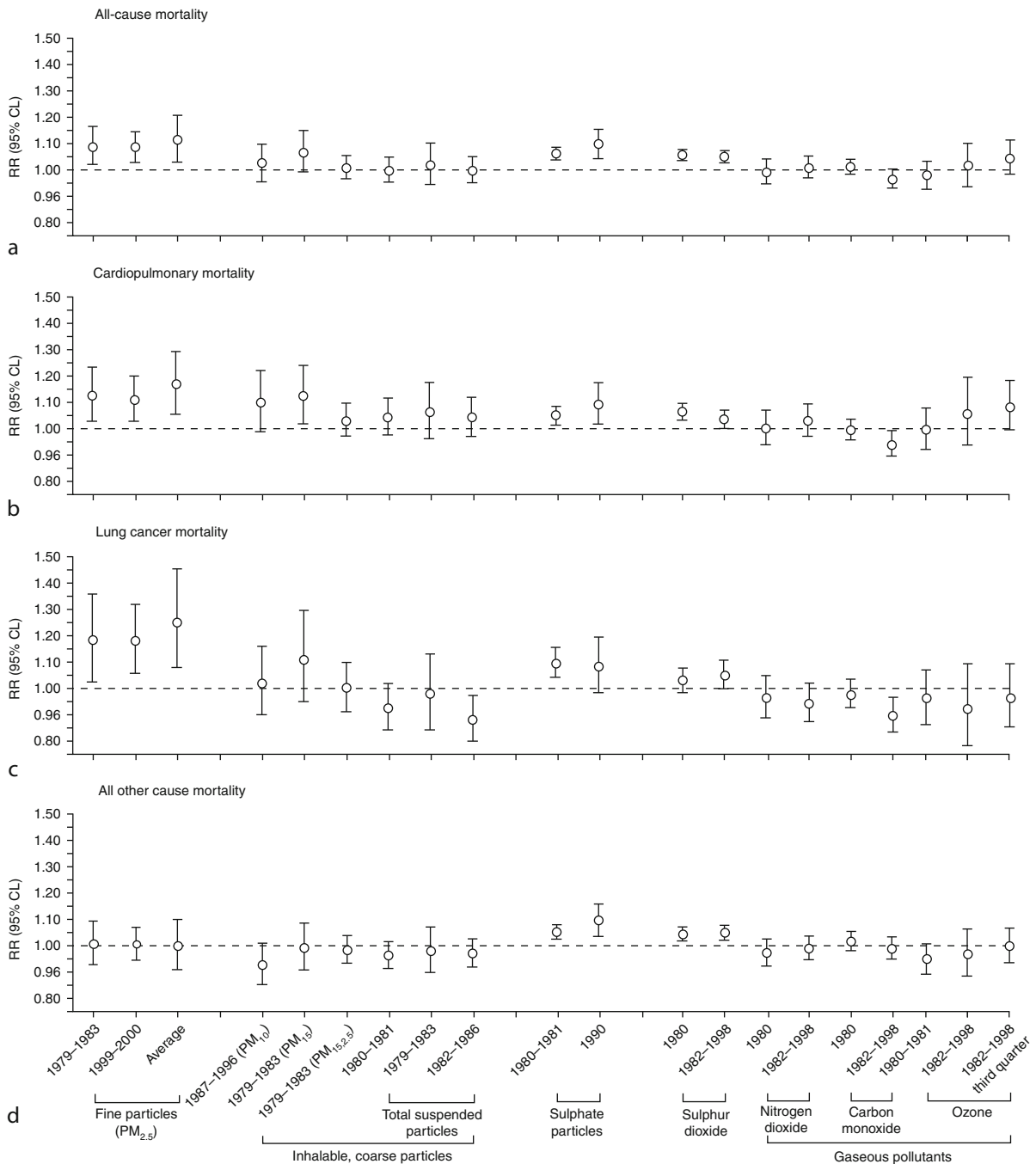
The Particulate Matter (PM) NAAQS was initially set in 1971 using Total Suspended Particulate (TSP) Matter as an indicator [36, 37, 39, 42, 64]. Later, the standard was revised with the TSP indicator replaced by  $PM_{10}$ .  $PM_{10}$  is defined as the size fraction collected using a device that includes 50% of the particle with an aerodynamic diameter of 10  $\mu m$ . A progressively smaller fraction of PM mass is collected as the particle size increases above 10  $\mu m$  and a progressively larger fraction of PM mass is sampled as particle size decreases below 10  $\mu m$ . Later, the standard was again changed with  $PM_{2.5}$  introduced as an indicator to complement the  $PM_{10}$  indicator.  $PM_{2.5}$ , the fine particle fraction indicator is defined as for  $PM_{10}$  indicator with a 50% cut point at 2.5  $\mu m$  aerodynamic diameter. To state the obvious, the  $PM_{2.5}$  fraction is contained within the  $PM_{10}$  fraction which is contained within the TSP sample. In recent years, increased attention has been focused on the ultrafine particle number which is contained within the  $PM_{2.5}$  size range. There has also been interest in characterizing health effects of the

$PM_{10-2.5}$  fraction which has been called the coarse particle fraction. Consideration is given to regulating this fraction with a separate NAAQS.

A detailed review of the evidence for health effects of PM is beyond the scope of this entry. Suffice it to say, abundant data are available on the health effects of PM, characterized using different PM indicators, from epidemiological and laboratory animal studies using a range of indicators of morbidity and mortality [65, 68, 84]. In setting the PM standards, the greatest weight, appropriately, has been given to the epidemiological evidence and, specifically, to those studies which can provide quantitative exposure-response information.

Two large studies have yielded especially valuable information that has played a key role in the setting of the PM NAAQS using the  $PM_{10}$  and  $PM_{2.5}$  indicators. The first of these is the Harvard Six Cities study designed by Ferris and colleagues [20, 23] to specifically evaluate the health effects of criteria air pollutants. In my opinion, this is one of the best studies ever conducted of the influence of air quality on health. It involves a 14–16-year follow-up of 8,000 subjects living in six communities in the United States selected to provide a gradient in several indices of air quality. Extensive efforts were made to characterize air quality. A second air quality study made opportunistic use of data from a population of 500,000 individuals who self-enrolled in 1982–1989 in an American Cancer Society Study [54]. A subgroup of 240,000 of these individuals lived in 50 metropolitan areas in which  $PM_{2.5}$  measurements were made for regulatory compliance purposes.

The vast amount of information available from the Six Cities and ACS studies has been reviewed by the EPA [68, 70, 84]. Selected data will be presented here to illustrate some of the key results that have influenced the setting of the PM NAAQS. Figure 7 summarizes an impressive array of information from the ACS study on all-cause mortality, cardiopulmonary mortality, lung cancer mortality, and all other cause mortality [54]. With control for smoking, education, and national status, the controlled forward stepwise inclusion of additional covariants had little influence on the estimated associations between fine particulate ( $PM_{2.5}$  indicator) air pollution on cardiopulmonary and lung cancer mortality. As may be noted, the inclusion of



**Air Quality Guidelines and Standards. Figure 7**

Relative risk (RR) ratio for adjusted mortality evaluated at subject-weighted mean concentrations. *CI* confidence interval,  $PM_{2.5}$  particles measuring less than  $2.5\ \mu m$  in diameter,  $PM_{10}$  particles measuring less than  $10\ \mu m$  in diameter,  $PM_{15}$  particles measuring less than  $15\ \mu m$  in diameter,  $PM_{2.5-15}$  particles measuring between  $2.5$  and  $15\ \mu m$  in diameter (Adapted from Pope et al. with permission)

cigarette smoking attenuated the estimated relative risk of  $PM_{2.5}$ , which is not surprising. When this is done it is necessary to normalize the smoking history to pack-years, for example, 1 pack/day for 25 years or 2 packs/day for 12.5 years equals 25 pack-years. The relative risks for an average current smoker (men and women combined, 22 cigarettes per day for 33.5 years with smoking initiated before age 18 years) (equivalent to 37 pack-years) were equal to 2.58, 2.89, and 14.80 for all-cause, cardiopulmonary, and lung cancer mortality, respectively. The magnitude of these relative risks is substantially greater than that of  $PM_{2.5}$  and emphasizes the critical importance of considering cigarette smoking in any study of air pollution. Moreover, it is important that investigators follow the lead of the ACS investigators and report the results of the analyses done for cigarette smoking at the same time they report relative risks for various air pollutants. These results on cigarette smoking provide valuable perspective for considering the results of analyses for the various air pollution indicators. A comparison of the relative risk estimates they obtain for smokers with those of other investigators serves as a “reality check” on their analyses.

The results from the Six Cities study and ACS-based studies are generally similar [28, 53]. Summary information from the two studies is given in Table 4 with a comparison made of the risk ratios for the most polluted versus least polluted cities. For comparison, results are also shown for typical smokers. Note that in this case the typical smoker has been defined as about a 25 pack-year smoker, about two thirds that of the current smoker discussed earlier for the ACS study.

Interestingly, the relative risk, for example, for lung cancer mortality is about two thirds of the value cited earlier, 9.73 versus 14.80. The range of the confidence intervals for the association between  $PM_{2.5}$  and cardiopulmonary mortality, even for these studies with thousands of subjects, emphasizes the challenges faced in teasing out small effects of air pollution in developed countries.

Laden et al. [29] extended the observations for the Six Cities Study Cohort by including deaths for 1990–1998 and compared those observations to the earlier time period (1974–1984) and for the entire period. The estimated city-specific average  $PM_{2.5}$  levels were lower in the extended follow-up period in the 1990s than in 1974–1989 and the mortality risk ratios were lower.

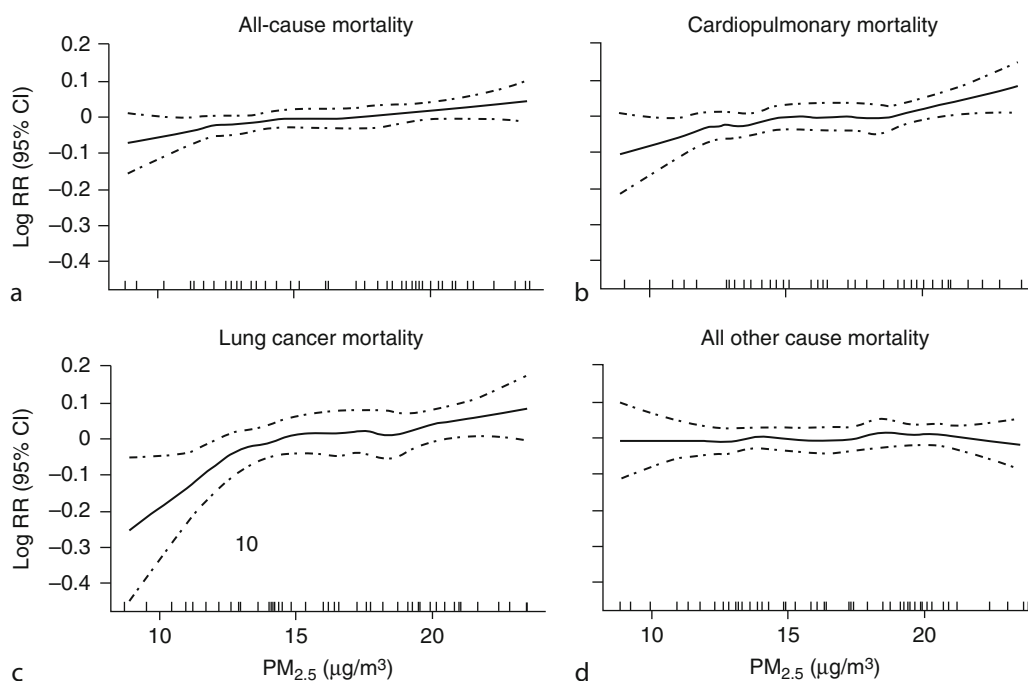
A critical issue in interpreting the data shown in Fig. 8 is the nature of the exposure-response relationship. It must be noted that this is the ambient concentration for PM that is being evaluated for an association, not PM exposure. The PM exposure cannot be evaluated since the estimated indoor PM concentrations cannot even be estimated. Pope et al. [54], using the ACS cohort data, evaluated the nonparametric smoothed exposure-response relationships between cause-specific and average  $PM_{2.5}$  (Fig. 8) [54]. They observed that the log-relative risk for all-cause, cardiopulmonary, and lung cancer mortality increased across the gradient of  $PM_{2.5}$  concentrations. Goodness-of-fit tests indicated that the associations were not significantly different from linear associations. The mean concentration of  $PM_{2.5}$  across the ACS cohort was  $14 \mu g/m^3$ .

**Air Quality Guidelines and Standards. Table 4** Comparison of mortality risk ratios (and 95% C.I.) for smoking and air pollution from the six cities and American Chemical Society (ACS) Prospective Cohort Studies

| Cause of death  | Current smoker <sup>a</sup> |                  | Particulate air pollution (most vs least polluted city) |                    |                  |
|-----------------|-----------------------------|------------------|---------------------------------------------------------|--------------------|------------------|
|                 | Six cities                  | ACS              | Six cities ( $PM_{2.5}$ )                               | ACS ( $PM_{2.5}$ ) | ACS (504)        |
| All             | 2.00 (1.51–2.65)            | 2.07 (1.75–2.43) | 1.25 (1.08–1.47)                                        | 1.17 (1.09–1.26)   | 1.15 (1.09–1.22) |
| Cardiopulmonary | 2.30 (1.56–3.41)            | 2.28 (1.79–2.91) | 1.37 (1.11–1.68)                                        | 1.31 (1.17–1.46)   | 1.26 (1.16–1.37) |
| Lung cancer     | 8.00 (2.97–21.6)            | 9.73 (5.96–15.9) | 1.37 (0.81–2.31)                                        | 1.03 (0.80–1.33)   | 1.36 (1.11–1.66) |
| All others      | 1.46 (0.89–2.39)            | 1.54 (1.19–1.99) | 1.01 (0.79–1.30)                                        | 1.07 (0.92–1.24)   | 1.01 (0.92–1.11) |

<sup>a</sup>Risk ratios for current cigarette smokers with approximately 25 pack-years (about average at enrollment for both studies) compared with never smokers

Source: Adapted from Pope and Dockery [53]



**Air Quality Guidelines and Standards. Figure 8**

Nonparametric smoothed exposure-response relationship. *Vertical lines* along the x-axes indicate a rug or frequency plot of mean fine particulate pollution. *CI* confidence interval,  $PM_{2.5}$ , fine particles measuring less than 2.5  $\mu m$  in diameter, *RR* relative risk (Adapted from Pope et al. with permission)

Enstrom [22] evaluated the relationship between fine particulate air pollution and total mortality among nearly 50,000 elderly Californians for 1973–2002 using a subpopulation of the ACS enrollees. He concluded that the results did not support a current relationship between fine particulate pollution and total mortality, although he could not rule out a small effect, particularly before 1983. The 2006 EPA analysis used to revise the PM NAAQS interpreted the available epidemiological data as not clearly supporting or refuting the existence of a threshold. The EPA analysis did calculate added risks down to an assumed threshold of 7.5  $\mu g/m^3$ ; a level on the order of the ambient  $PM_{2.5}$  concentrations found in many US cities.

While the estimated or excess risk for the various PM indicators are quite small, it is apparent that when used in company with standard baseline data for specific cities with populations measured in the millions, the calculated number of excess deaths appears large. The USEPA Administrator can consider such quantitative risk assessment results in making a “judgment call”

in setting the standards. However, the Administrator is explicitly prohibited from considering cost when setting the standard. In separate regulatory impact analyses, the risk assessment results are monetized in economic analyses used to estimate the benefits associated with achieving a revised standard compared to maintaining an existing standard. The regulatory impact analysis prepared for ozone and particulate matter [67] and separately for particulate matter [76] and ozone [78] are examples.

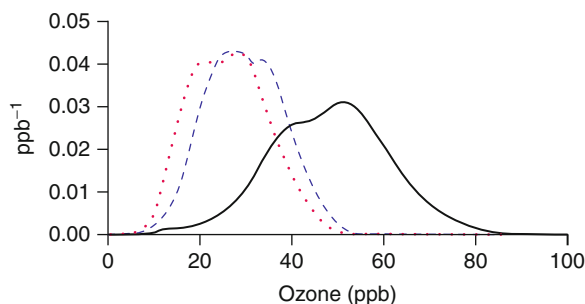
Substantial controversy developed in the final stages of revision of the  $PM_{2.5}$  standard announced in 2006. A majority of the EPA’s Clean Air Scientific Advisory Committee’s (CASAC) PM Panel recommended that the EPA Administrator reduce the  $PM_{2.5}$  24-h averaging time standard from the previous level of 65  $\mu g/m^3$  to a level in the range of 25–35  $\mu g/m^3$  and that for the  $PM_{2.5}$  annual averaging time the standard be reduced from the previous level of 15  $\mu g/m^3$  to a level in the range of 12–14  $\mu g/m^3$ . I was in the minority and argued that identifying a specific bright line, upper bound such

as  $14 \mu\text{g}/\text{m}^3$  was an issue that went beyond scientific interpretation of data and was a policy judgment as to acceptable risk. The Administrator, ultimately, revised the  $\text{PM}_{2.5}$  NAAQS, 24-h averaging time standard, to  $35 \mu\text{g}/\text{m}^3$  and retained the annual standard at  $15 \mu\text{g}/\text{m}^3$  [74, 75]. Obviously, the majority of CASAC was not happy with the Administrator's policy judgment call.

## Ozone

Ozone is an interesting air pollutant because it is not directly emitted from industrial sources except in rare circumstances. The majority of Ozone found in the ambient environment is formed from reactions of Volatile Organic Compounds (VOC) and Nitrogen Oxides ( $\text{NO}_x$ ) in the presence of sunlight. The VOCs originate both from natural and industrial sources and the  $\text{NO}_x$  is largely from combustion of hydrocarbon fuels. Ozone concentrations typically have distinct diurnal cycles with the highest levels observed at mid-day and the lowest concentrations at night in the summer. Lower concentrations are typically observed in the winter than in the summer. This is the case because of the manner in which ozone is formed with sunlight driving the reactions between  $\text{NO}_x$  and VOCs.

To provide context for the discussion of the Ozone NAAQS it is important to understand the background levels of ambient ozone. A team of investigators at Harvard University led by Daniel Jacob have made major contributions to understanding tropospheric ambient ozone levels around the world using a global chemical transport model (GEOS-Chem). The Harvard team [90] reported the use of the GEOS-Chem model with  $1^\circ$  latitude by  $1^\circ$  longitude spatial resolution to quantify the effects of anthropogenic emissions from Canada, Mexico, and outside North America to estimate daily maximum 8-h average ozone concentrations in US surface air. Their simulations for summer 2001 estimated mean North American and US background contributions of  $26 \pm 8$  ppb and  $30 \pm 8$  ppb, as obtained by eliminating anthropogenic emissions in North America versus in the United States only. The background 8-h highest Ozone concentrations approached 60 ppb. The frequency distribution of the simulated daily 8-h-maximum surface Ozone concentrations are shown in Fig. 9. It is important to recognize that the statistical form of the Ozone NAAQS is



**Air Quality Guidelines and Standards. Figure 9** Frequency distribution of the simulated daily 8-h-maximum surface ozone (black solid line). North American background (red dotted line) and US background (blue dashed line), for the contiguous United States during June–August 2001. The frequency distributions are constructed from the  $1^\circ \times 1^\circ$  daily model output (From [90])

directed at minimizing the occurrence of the highest ozone events, the right side tail of the distributions in the figure. The high levels of background Ozone will make it challenging to achieve any Ozone NAAQS set in the range of 60–80 ppb (8-h daily maximum).

The initial Ozone NAAQS in the United States was set in 1971 using photochemical oxidants as an indicator and with a 1-h averaging time [7, 64]. The standard was later revised with ozone as the indicator. Later the averaging time was changed from a 1-h maximum during a 24-h period to an 8-h rolling average maximum.

There is an abundant database on the health effects of ozone from epidemiological investigations and toxicity studies conducted in laboratory animals [38, 71, 80, 81]. In addition, there is a substantial amount of data acquired from studies conducted with human volunteers exposed to controlled levels of ozone for short periods of time. In the setting of the Ozone NAAQS, quantitative exposure-response data from human studies have been given the greatest weight. This is especially appropriate since there are well-established quantitative differences between laboratory animal species and humans in the disposition of inhaled ozone which minimize the utility of the data from laboratory animals for quantitatively estimating human risk of ozone exposures.

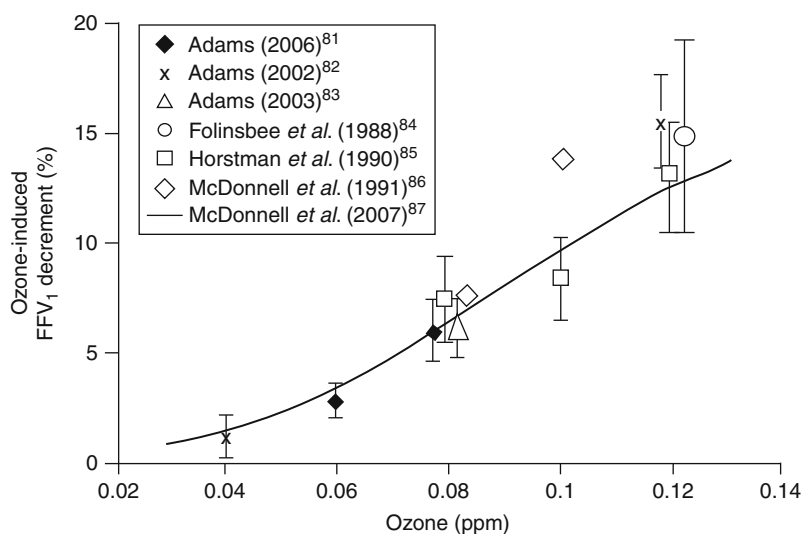


The effects of ozone on the respiratory tract are well known, thus it is not surprising that consideration of these effects has had a primary influence on the setting of the Ozone NAAQS. In the review culminating with a revised standard in March 2008, the information from pulmonary function evaluations conducted on young adult volunteers undergoing moderate exercise while exposed to various concentrations of ozone for 6.6 h played a prominent role. The key parameter evaluated, changes in the forced expiratory volume in 1 s ( $FEV_1$ ), are shown in Fig. 10 [1–3, 13, 24, 26, 44, 45]. Data from individuals exposed to 0.08 ppm ozone and higher were considered in the previous review of the ozone standard. In the most recent review, data were available from the studies of Adams and colleagues [3] with 0.06 and 0.04 ppm exposures conducted to complement earlier studies conducted with 0.08 and 0.012 ppm, Adams [3] interpreted his studies as showing a statistically significant effect at 0.08 ppm, effectively reproducing the effects seen in earlier studies. However, he interpreted his studies at 0.06 and 0.04 ppm as not showing statistically significant changes compared to exposure to clean air. The USEPA reanalyzed the data of Adams and colleagues using different statistical procedures [13]. The USEPA reanalysis was included in the final compilation of

information provided to the EPA Administrator to render a judgment on the revision of the Ozone NAAQS.

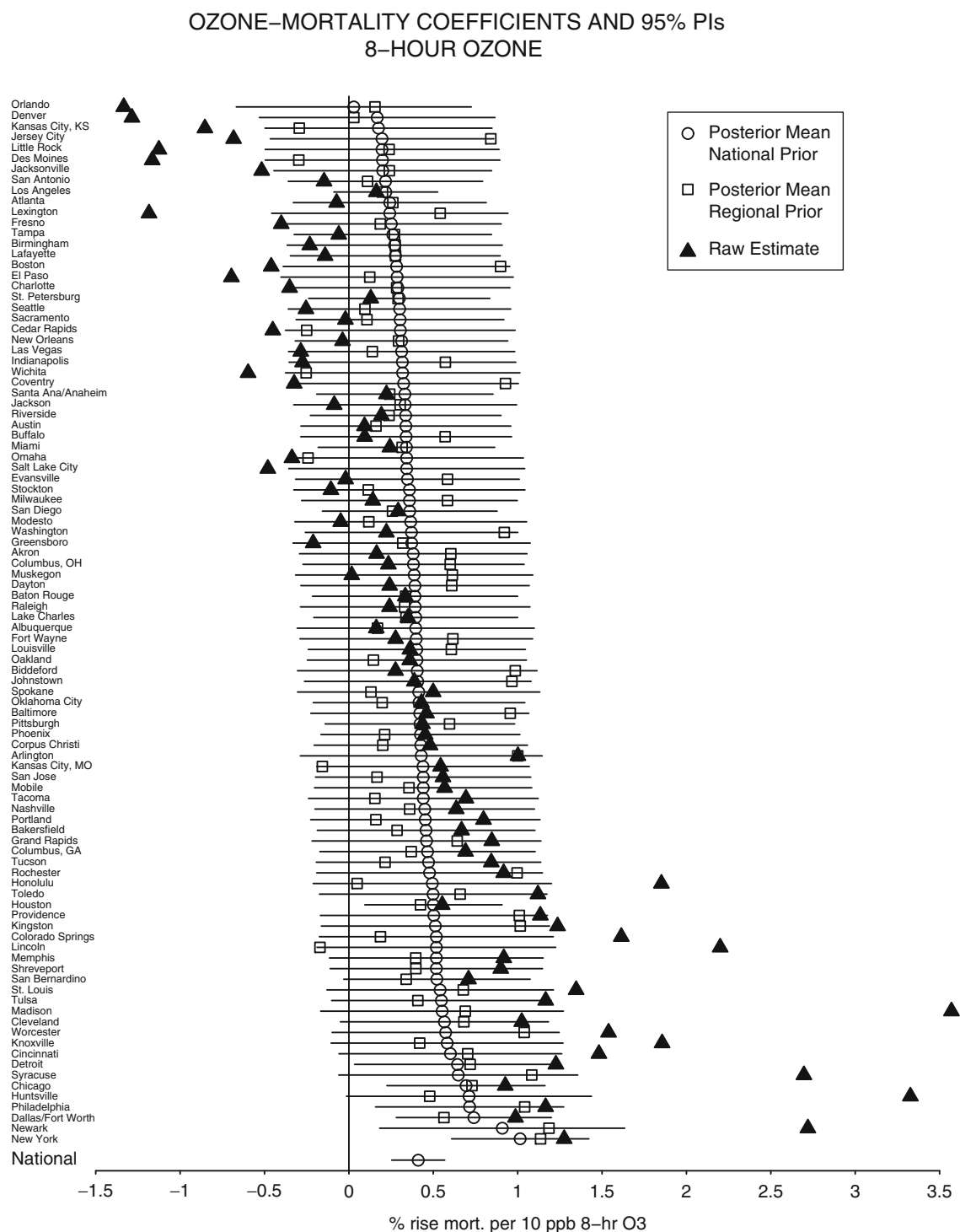
A key piece of information on the health effects of ozone came from the time-series study of Bell et al. [8, 9]. Their analysis used a 24-h averaging time and, thus, was not directly applicable to setting the Ozone NAAQS with an 8-h averaging time. Smith et al. [58] extended the analyses of Bell et al. [8] focusing on the development of city-specific ozone mortality estimates. The analyses of both Bell et al. [8] and Smith et al. [58] used the National Air Pollution Morbidity and Mortality air pollution database assembled at Johns Hopkins University. The results for the 8-h averaging time are shown in Fig. 11. The degree of intercity heterogeneity is noteworthy including recognition that only 5 cities of the 98 cities evaluated had statistically significant effects of ozone on acute mortality.

Smith et al. [58] called attention to this intercity heterogeneity across the United States (Fig. 12). The ozone effects are most striking in the Northeastern United States and absent in the Western United States including the Los Angeles Basin. It is not surprising that Smith et al. [58] caution against the use of national estimates of ozone concentration-response functions.



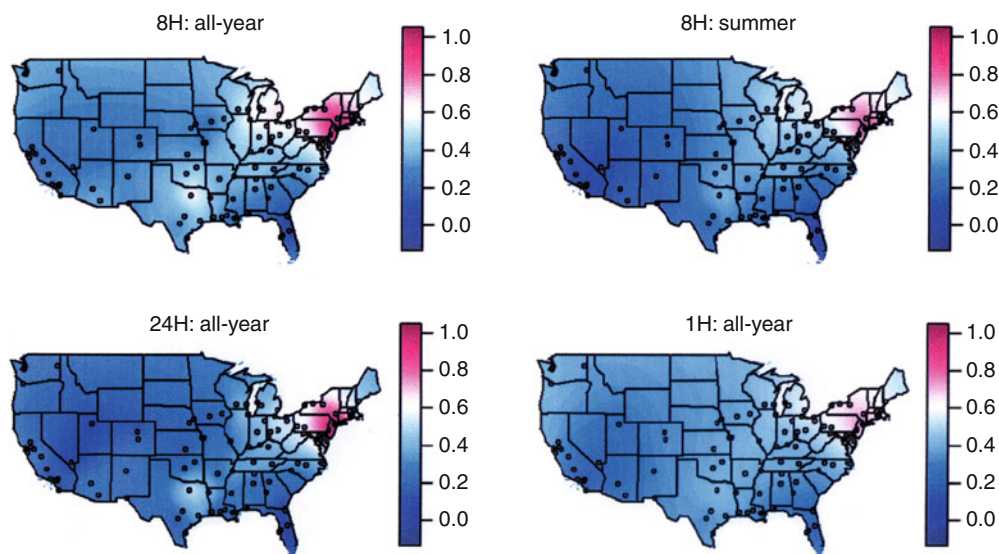
**Air Quality Guidelines and Standards. Figure 10**

Changes in the lung function of young health adults following a 6.6-h exposure to ozone.  $FEV_1$  forced expiratory volume in 1 s, ppm parts per million (Adapted from Brown et al. with permission)



**Air Quality Guidelines and Standards. Figure 11**

Ninety-five percent posterior intervals for the ozone-mortality coefficients, based on 8-h ozone, all-year data. The Bayesian posterior estimates under the “national prior” (circles) are shown alongside those for the “regional prior” (squares) and the raw maximum likelihood estimates (triangles) (Adapted from [58])



**Air Quality Guidelines and Standards. Figure 12**

Map of spatially dependent ozone-mortality coefficient for 8-h ozone (all-year data), 8-h ozone (summer data), 24-h ozone (all-year data), and 1-h ozone (all-year data) (Adapted from [58])

The EPA's CASAC recommended to the Administrator that he revise the Ozone NAAQS with an 8-h averaging time, from 0.08 ppm to a level within the range of 0.060–0.070 ppm. Ultimately, the Administrator exercised his judgment in revising the standard to 0.075 ppm. The CASAC Ozone Panel expressed its displeasure at the Administrator's setting the NAAQS at a level higher than they recommended. In my opinion, the science alone can never be sufficient for establishing a specific numerical standard. The setting of the standard also involves judgment as to acceptable risk, a judgment call I viewed as being delegated to the Administrator, not the Advisory Committee, by the Clean Air Act. After the revised standard was promulgated, the EPA's statistical reanalysis [13] was published with a conclusion that the Adams [3] data demonstrated a statistically significant, but small biological effect at 0.060 ppm. These published findings as well as any new data that may become available from studies conducted with human volunteers exposed to ozone at concentrations of less than 0.080 ppm will certainly be considered in the current review (post-2008) and potential revision of the Ozone NAAQS.

The setting of both the PM and Ozone NAAQS serves to illustrate the complex interplay between

science and policy judgment in the setting of guidelines and standards [41]. In my opinion, the selection of any specific numerical standard for exposure to a chemical involves both science and a policy judgment as to an acceptable risk. The scientific information can identify for a particular level of exposure an associated level of risk (including an absence of excess risk) and the associated uncertainty in the estimate. However, the selection of a specific level for the standard goes beyond the science and represents a policy judgment or decision. The issue of the role of science and judgment in standard setting is drawn to the forefront by agents that do not have clear thresholds in their exposure-response relationships. In these cases, some level of excess risk may be calculated for exposures that extend down to ambient levels routinely observed in developed countries.

In March 2008, then EPA Administrator Stephen Johnson announced a revised Ozone NAAQS (8-h averaging time) of 75 ppb, a reduction from the previous level of 0.08 ppm (which by rounding convention was effectively 84 ppb) set in 1997 [81]. In setting the 75 ppb level, he explicitly noted that he was not explicitly following the advice of the CASAC which advised that he set the standard in the range of 60–70 ppb.

He noted that he viewed the CASAC advice as a blend of science and policy judgment which he was not compelled to follow. The previous statistical form for the standard was not changed.

The Administrator also elected to set the Secondary (Welfare) Standard equal to the Primary (Health) Standard. The 1997 Secondary Standards for Ozone [65] had been set equal to the Primary Standard based on the view that attainment of the Primary Standard would be protective of welfare impacts and effects on plants. The CASAC had recommended that the averaging time form of the Secondary Standard be changed to a more complex form that used a weighted average of the Ozone concentration over a growing season. Thus, it was not as easy to compare the Primary and Secondary Standards. As an aside, the revision of the Ozone NAAQS in March 2008 also set the dates for completing the next regularly scheduled 5-year review, March 2013.

The failure of the EPA Administrator to explicitly follow the advice of CASAC was greeted with outrage by the CASAC Ozone Panel. The Committee took the unusual step, post-final announcement of the new Ozone NAAQS, of holding a Panel meeting to discuss the new Standard and writing a letter to the Administrator reminding him of their previous advice and calling for the Administrator to follow their advice in future revisions of the standard [41].

In January 2009, President Barak Obama took office and Lisa Jackson was appointed EPA Administrator. Coincident with Obama taking office his Chief of Staff, Rahm Emanuel, took the customary step of reminding newly designated Senior Agency officials of the opportunity to review and, perhaps, revise any rules and regulations that originated with the prior Administration and not yet finalized [21]. Administrator Lisa Jackson took the bold step of announcing that she was going to reconsider the Ozone NAAQS set by Administrator Johnson in March 2008. In taking this action she was in essence saying – if I had been in office as EPA Administrator earlier I would have made a different policy decision. Administrator Jackson proceeded to publish a proposal to “reconsider” the Ozone NAAQS using the same science used as a basis of the March 2008 standard [81]. The proposal indicated Administrator Jackson intended to follow the advice of

CASAC and set a primary standard (8-h averaging time) in the range of 60–70 ppb. The proposed date for release of a final standard was announced, a target date that was revised on multiple occasions. A Presidential Executive Order issued by President Clinton [18] requires that any proposed federal rule or action that will result in annual costs in excess of \$100 million per year must be reviewed by the Office of Information and Regulatory Affairs in the Office of Management and Budget. On September 2, 2011, President Obama announced in a press release that he had advised EPA Administrator Jackson to not proceed with the “reconsideration” of the Ozone NAAQS [91]. At the same time, a brief summary of the conclusions of the OIRA review was released by the Director of that Office, Cass Sunstein [62]. It briefly noted that the EPA “reconsideration” proposal (a) was discretionary and not required by the CAA 5-year review schedule, (b) was not based on the best available science (the record was based on scientific literature that had been published in most cases before 2006), and (c) was not consistent with the President’s desire for regulations that reduced uncertainty.

Environment groups greeted the President’s announcement with dismay, while Business applauded it. A Forum ([Forum.HSPH.org](http://Forum.HSPH.org)) held on September 21, 2011 at the Harvard School of Public Health entitled “Smog or Jobs: The Impact of Tighter Ozone Control” provides insight into the controversial nature of President Obama’s decision informed by the views of his advisor, Sunstein. The current review of the Ozone NAAQS initiated in 2008 will now proceed. A key initial step in the review process occurred with release in March 2011 of new integrated science assessment for ozone [88]. A proposal for the Ozone Standard should be issued in 2012 for public comment and a final rule for continuing or revising the Ozone NAAQS should be issued in March 2013 if EPA adheres to the 5-year schedule specified in the Clean Air Act [16].

### Current Air Quality Standards and Guidelines

As noted earlier, the National Ambient Air Quality Standards established by the US Environmental Protection Agency are constantly changing. Indeed, the Clean Air Act specifies that each NAAQS should be reviewed

**Air Quality Guidelines and Standards. Table 5** Summary information on criteria air pollutants regulated as NAAQS<sup>a</sup> by the USEPA

| Pollutant          | Sources                                                                                                                                             | Key effects of concern                                                                     | Subpopulations of concern                                                       |
|--------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Ozone              | Photochemical oxidation of nitrogen oxides (primarily from combustion) and volatile organic compounds (from stationary and mobile sources)          | Decreased pulmonary function, lung inflammation, increased respiratory hospital admissions | Children, people with preexisting lung disease outdoor-exercising health people |
| Nitrogen dioxide   | Photochemical oxidation of nitric oxide (primarily from combustion of fossil fuels) and direct emissions (primarily from combustion of natural gas) | Respiratory illness, decreased pulmonary function                                          | Children, people with preexisting lung disease                                  |
| Sulfur dioxide     | Primarily combustion of sulfur-containing fossil fuels; also smelters, refineries, and others                                                       | Respiratory injury or death, decreased pulmonary function                                  | Children, people with preexisting lung disease (especially asthma)              |
| Particulate matter | Direct emission of particles during combustion, industrial processes, gas reactions and condensation and coagulation natural sources                | Injury and death                                                                           | Children, people with preexisting heart and lung disease                        |
| Carbon monoxide    | Combustion of fuels, especially by mobile sources                                                                                                   | Shortening of time to onset angina and other heart effects                                 | People with coronary artery disease                                             |
| Lead               | Leaded gasoline (prior to phase-out in gasoline); point sources such as Pb mines, smelters, and recycling operations                                | Developmental neurotoxicity                                                                | Children                                                                        |

<sup>a</sup>NAAQSs are specified by indicator concentration, averaging time, and statistical form. The latter used to determine compliance and is too complex for coverage here. However, suffice it to note that both the specific numerical level *and* the statistical form determine the stringency of the standard

every 5 years. Although this schedule has been rarely met, the reviews continue and changes in the NAAQS have occurred. There is no specified schedule for review of the WHO Air Quality Guidelines. However, it would be reasonable to assume that they will be periodically reviewed and possibly revised as new information is required.

The NAAQS as of September 18, 2011 are shown in [Tables 5](#) and [6](#) in summary form. The reader interested in the details that support each of the NAAQS is referred to the US EPA's Website for criteria air pollutants where detailed documents can be found on each of the Criteria Pollutants. This includes information on previous reviews as well as reviews that are currently in progress.

The final rule for each of the primary (health) NAAQS published in the Federal Register are as follows: particulate matter [\[75\]](#), Lead [\[73\]](#), Ozone [\[82\]](#),

Nitrogen dioxide [\[86\]](#), Sulfur dioxide [\[87\]](#), and Carbon monoxide [\[89\]](#). Each of these final rules contains references to the Criteria Document (or Integrated Science Assessment), Risk and exposure assessment, and proposed rule that undergirds the final rule. In addition, the EPA Web site for criteria pollutants provides linkages to the Regulatory Impact Analyses (RIA) that has been conducted on the proposed final rule. The RIA contain information on the monetized benefits of the rule and the associated costs of achieving the NAAQS. In accordance with the US Supreme Court ruling in *American Trucking Associations v. EPA* [\[4\]](#) and *Whitman v. American Trucking Associations* [\[93\]](#), the costs of implementing the NAAQS may not be considered by the EPA Administrator in the setting of the NAAQS. However, these costs can be considered in developing a schedule for implementation of each NAAQS.



**Air Quality Guidelines and Standards. Table 6** National Ambient Air Quality Standards for criteria pollutants

| Pollutant                               | Primary standards                   |                                          | Secondary standards |                  |
|-----------------------------------------|-------------------------------------|------------------------------------------|---------------------|------------------|
|                                         | Level                               | Averaging time                           | Level               | Averaging time   |
| Carbon monoxide                         | 9 ppm (10 mg/m <sup>3</sup> )       | 8-h <sup>a</sup>                         | None                |                  |
|                                         | 35 ppm (40 mg/m <sup>3</sup> )      | 1-h <sup>a</sup>                         |                     |                  |
| Lead                                    | 0.15 µg/m <sup>3</sup> <sup>b</sup> | Rolling 3-month average                  | Same as primary     |                  |
| Nitrogen dioxide                        | 53 ppb <sup>c</sup>                 | Annual (arithmetic average)              | Same as primary     |                  |
|                                         | 100 ppb                             | 1-h <sup>d</sup>                         | None <sup>i</sup>   |                  |
| Particulate matter (PM <sub>10</sub> )  | 150 µg/m <sup>3</sup>               | 24-h <sup>e</sup>                        | Same as primary     |                  |
| Particulate matter (PM <sub>2.5</sub> ) | 15.0 µg/m <sup>3</sup>              | Annual <sup>f</sup> (arithmetic average) | Same as primary     |                  |
|                                         | 35 µg/m <sup>3</sup>                | 24-h <sup>g</sup>                        |                     |                  |
| Ozone                                   | 0.075 ppm (2008 std)                | 8-h <sup>h</sup>                         | Same as primary     |                  |
|                                         | 0.08 ppm (1997 std)                 | 8-h <sup>i</sup>                         |                     |                  |
|                                         | 0.12 ppm                            | 1-h <sup>j</sup>                         |                     |                  |
| Sulfur dioxide                          | 0.03 ppm                            | Annual (arithmetic average)              | 0.5 ppm             | 3-h <sup>a</sup> |
|                                         | 0.14 ppm                            | 24-h <sup>a</sup>                        |                     |                  |
|                                         | 75 ppb <sup>k</sup>                 | 1-h                                      | None <sup>l</sup>   |                  |

<sup>a</sup>Not to be exceeded more than once per year

<sup>b</sup>Final rule signed October 15, 2008. The 1978 lead standard (1.5 µg/m<sup>3</sup> as a quarterly average) remains in effect until 1 year after an area is designated for the 2008 standard, except that in areas designated nonattainment for the 1978 standard, the 1978 standard remains in effect until implementation plans to attain or maintain the 2008 standard are approved

<sup>c</sup>The official level of the annual NO<sub>2</sub> standard is 0.053 ppm, equal to 53 ppb, which is shown here for the purpose of clearer comparison to the 1-h standard

<sup>d</sup>To attain this standard, the 3-year average of the 98th percentile of the daily maximum 1-h average at each monitor within an area must not exceed 100 ppb (effective January 22, 2010)

<sup>e</sup>Not to be exceeded more than once per year on average over 3 years

<sup>f</sup>To attain this standard, the 3-year average of the weighted annual mean PM<sub>2.5</sub> concentrations from single or multiple community-oriented monitors must not exceed 15.0 µg/m<sup>3</sup>

<sup>g</sup>To attain this standard, the 3-year average of the 98th percentile of 24-h concentrations at each population-oriented monitor within an area must not exceed 35 µg/m<sup>3</sup> (effective December 17, 2006)

<sup>h</sup>To attain this standard, the 3-year average of the fourth-highest daily maximum 8-h average ozone concentrations measured at each monitor within an area over each year must not exceed 0.075 ppm (effective May 27, 2008)

<sup>i</sup>(a) To attain this standard, the 3-year average of the fourth-highest daily maximum 8-h average ozone concentrations measured at each monitor within an area over each year must not exceed 0.08 ppm.

(b) The 1997 standard – and the implementation rules for that standard – will remain in place for implementation purposes as EPA undertakes rule making to address the transition from the 1997 ozone standard to the 2008 ozone standard.

(c) EPA is in the process of reconsidering these standards (set in March 2008)

<sup>j</sup>(a) EPA revoked the 1-h ozone standard in all areas, although some areas have continuing obligations under that standard (“anti-backsliding”).

(b) The standard is attained when the expected number of days per calendar year with maximum hourly average concentrations above 0.12 ppm is ≤1

<sup>k</sup>Final rule signed June 2, 2010. To attain this standard, the 3-year average of the 99th percentile of the daily maximum 1-h average at each monitor within an area must not exceed 75 ppb

<sup>l</sup>On July 12, 2011, the EPA Administrator proposed to set Secondary NAAQS for NO<sub>2</sub> and SO<sub>2</sub> for a 1-h averaging time at concentrations equal to the Primary Standards, NO<sub>2</sub> – 100 ppb and SO<sub>2</sub> – 75 ppb (Note: This footnote was added by Roger O. McClellan)

Source: Taken from US EPA Website, July 26, 2011

**Air Quality Guidelines and Standards. Table 7** WHO air quality guidelines and interim targets for particulate matter<sup>a</sup>

| Annual mean concentration                           |                                          |                                           |                                                                                                                                                                                                        |
|-----------------------------------------------------|------------------------------------------|-------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                     | PM <sub>10</sub><br>(µg/m <sup>3</sup> ) | PM <sub>2.5</sub><br>(µg/m <sup>3</sup> ) | Basis for the selected level                                                                                                                                                                           |
| Interim Target-1 (IT-1)                             | 70                                       | 35                                        | These levels are associated with about a 15% high long-term mortality risk relative to the AQG level                                                                                                   |
| Interim Target-2 (IT-2)                             | 50                                       | 25                                        | In addition to other health benefits, these levels lower the risk of premature mortality by approximately 6% (2–11%) relative to the IT-1 level                                                        |
| Interim Target-3 (IT-3)                             | 30                                       | 15                                        | In addition to other health benefits, these levels reduce the mortality risk by approximately 6% (2–11%) relative to the IT-2 level                                                                    |
| Air Quality Guideline (AQG)                         | 20                                       | 10                                        | These are the lowest levels at which total, cardiopulmonary and lung cancer mortality have been shown to increase with more than 95% confidence in response to long-term exposure to PM <sub>2.5</sub> |
| 24-h concentrations [99th percentile (3 days/year)] |                                          |                                           |                                                                                                                                                                                                        |
|                                                     | PM <sub>10</sub><br>(µg/m <sup>3</sup> ) | PM <sub>2.5</sub><br>(µg/m <sup>3</sup> ) | Basis for the selected level                                                                                                                                                                           |
| Interim Target-1 (IT-1)                             | 150                                      | 75                                        | Based on published risk coefficient from multi-center studies and meta-analyses (about 5% increase of short-term mortality over the AQG value)                                                         |
| Interim Target-2 (IT-2)                             | 100                                      | 50                                        | Based on published risk coefficients from multiple-center studies and meta-analyses (about 2.5% increase of short-term mortality over the AQG value)                                                   |
| Interim Target-3 (IT-3) <sup>a</sup>                | 75                                       | 37.5                                      | Based on published risk coefficients from multi-center studies and meta-analyses (about 1.2% increase in short-term mortality over the AQG value)                                                      |
| Air quality                                         | 50                                       | 25                                        | Based on relationship between 24-h and annual PM levels been shown to increase with more than 95%                                                                                                      |

<sup>a</sup>The use of PM<sub>2.5</sub> guideline value is preferred

Source: WHO [96]

The current WHO air quality guidelines are summarized in [Tables 7, 8, 9, and 10](#). The basis for these guidelines is contained in WHO documents [95, 96].

### Future Directions

This entry has provided an overview of Air Quality Guidelines and Standards focusing on the standards developed by the US Environmental Protection Agency and the Guidelines developed by the World Health Organization. The latter Guidelines, although stated to provide guidance for use around the world, have focused on Europe. For Europe, the WHO Guidelines have provided valuable input to national and local authorities for developing legal binding air quality standards. The Standards developed by European nations and other nations round the world have not been reviewed in detail in this entry.

The section on “[Air Quality Management Framework](#)” emphasized that Air Quality Standards and Guidelines were a key element of the overall scheme; however, the Standards and Guidelines are just one element of an integrated air quality management scheme. It is noteworthy that substantial progress has been made in improving air quality in the United States and in western European countries. That progress has no doubt been facilitated by the Standards and Guidelines. However, setting up goal posts (Standards and Guidelines) does not necessarily mean the goals will be achieved. Specific emission reductions informed by the Standards and Guidelines are ultimately required to achieve the goals. In the United States, major regulations have mandated reductions in emissions of particulate matter, SO<sub>2</sub>, and NO<sub>2</sub> from stationary sources such as power plants and from mobile sources such as passenger cars and on-road trucks and, more recently,

**Air Quality Guidelines and Standards. Table 8** WHO air quality guideline and interim target for ozone: 8-h concentrations

|                             | Daily maximum 8-h mean ( $\mu\text{g}/\text{m}^3$ ) | Basis for selected level                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|-----------------------------|-----------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| High levels                 | 240                                                 | Significant health effects; substantial proportion of vulnerable populations affected                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Interim Target-1 (IT-1)     | 160                                                 | <p>Important health effects; does not provide adequate protection of public health</p> <p>Exposure to this level of ozone is associated with:</p> <ul style="list-style-type: none"> <li>• Physiological and inflammatory lung effects in healthy exercising young adults exposed for periods of 6.6 h</li> <li>• Health effects in children (based on various summer camp studies in which children were exposed to ambient ozone levels)</li> <li>• An estimated 3–5% increase in daily mortality<sup>a</sup> (based on findings of daily time-studies)</li> </ul>                                                                  |
| Air Quality Guideline (AQG) | 100                                                 | <p>Provides adequate protection of public health, though some health effects may occur below this level. Exposure to this level of ozone is associated with:</p> <ul style="list-style-type: none"> <li>• An estimated 1–2% increase in daily mortality<sup>a</sup> (based on findings of daily time-series studies)</li> <li>• Extrapolation from chamber and field studies based on the likelihood that real-life exposure tends to be repetitive and chamber studies exclude highly sensitive or clinically compromised subjects, or children</li> <li>• Likelihood that ambient ozone is a marker for related oxidants</li> </ul> |

<sup>a</sup>Deaths attributable to ozone. Time-series studies indicate an increase in daily mortality in the range of 0.3–0.5% for every 10  $\mu\text{g}/\text{m}^3$  increment in 8-h ozone concentrations above an estimated baseline level of 70  $\mu\text{g}/\text{m}^3$

**Air Quality Guidelines and Standards. Table 9** WHO air quality guidelines and interim targets for  $\text{SO}_x$ : 24-h and 10-min concentrations

|                                      | 24-h average ( $\mu\text{g}/\text{m}^3$ ) | 10-min average ( $\mu\text{g}/\text{m}^3$ ) | Basis for selected level                                                                                                                                                                                                                                                                                                                                                                           |
|--------------------------------------|-------------------------------------------|---------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Interim Target-1 (IT-1) <sup>a</sup> | 125                                       | –                                           |                                                                                                                                                                                                                                                                                                                                                                                                    |
| Interim Target-2 (IT-2)              |                                           |                                             | Intermediate goal based on controlling either motor vehicle emissions, industrial emissions, and/or emissions from power production. This would be a reasonable and feasible goal for some developing countries (it could be achieved within a few years) which would lead to significant health improvements that, in turn, would justify further improvements (such as aiming for the AQG value) |
| Air Quality Guideline (AQG)          | 20                                        | 500                                         |                                                                                                                                                                                                                                                                                                                                                                                                    |

<sup>a</sup>Formerly the WHO Air Quality Guideline [95]

from off-road equipment, rail locomotives, and ships. These regulations have resulted in substantial reductions in the emissions of the criteria pollutants or their

precursors while the economy also grew until recently. A stagnant economy in 2011 is raising questions as to the wisdom of further reductions in NAAQS

**Air Quality Guidelines and Standards. Table 10** WHO (European) guideline values for carbon monoxide

|                                    |        |
|------------------------------------|--------|
| 100 mg/m <sup>3</sup> <sup>a</sup> | 15 min |
| 60 mg/m <sup>3</sup> <sup>a</sup>  | 30 min |
| 30 mg/m <sup>3</sup> <sup>a</sup>  | 1 h    |
| 10 mg/m <sup>3</sup>               | 8 h    |

<sup>a</sup>Exposure at these levels should be no longer than the indicated times and should not be repeated within 8 h

Source: WHO [95]

and aggressive regulatory actions to further reduce emissions.

Since the Clean Air Act Amendments of 1970 were passed in the United States there has been a steady increase in research on the health effects of air pollutants. The increase in research has been most notable since the mid-1990s when major controversies arose over the health effects of PM<sub>2.5</sub>. Indeed, what has developed might be characterized as a “clean air enterprise” involving air quality regulators, the academic research community, and nongovernmental organization advocacy groups. Some individuals would quickly note that this enterprise is counterbalanced by corporations and industrial groups advocating for the status quo or “dirty air.” This polarized approach has frequently resulted in the focus being on a single risk factor, air quality, rather than the overall well-being of the health of societies around the world.

The earlier section on “[Context for Air Quality Impacts](#)” noted the multitude of risk factors that impact on health. That section contrasted the burden of disease attributed to the various risk factors including urban air pollution and indoor smoke from household use of solid fuels. The contrast between the burden of disease for low- and middle-income countries versus high-income countries was striking. In the former, 3.7% of deaths were associated with indoor smoke and 1.5% from urban air pollutions. In contrast, in the high-income countries urban air pollution was estimated to be associated with 1.0% of the deaths and no deaths were estimated to be associated with indoor smoke. The health hazards associated with indoor smoke from use of solid fuels for cooking and heating has not been reviewed in this entry. It is sufficient to note here that the greatest hazards from indoor

cooking are those related to burning of wood, agricultural residue, and coal in the same manner as was commonplace around the world centuries ago. It is clear that in the future a major challenge exists to improve the income of families in low- and middle-income countries so they can move away from uncontrolled combustion of solid fuels indoors. It is well recognized that improving the economy of a country or region inevitably requires increases in energy usage and the availability of clean water and adequate foodstuffs. A major challenge exists in providing adequate energy, in the form of distributed electricity and liquid fuels for transportation and also avoiding negative impacts on air quality. Major progress has been made in the highest income countries in the world in developing low-emission stationary sources for generating electrical power and low emission internal combustion [both spark ignition and compression ignition (diesel)] engines for on- and off-road use. The challenge is to facilitate the transfer of those low emission technologies to the rest of the world.

With the above as background, it is appropriate to return to the central focus of this entry – air quality standards and guidelines and their role in the future. The past 40 years have been a period of major activity in creating a science based on the health effects of air pollutants and the use of that science to inform the development of air quality standards and guidelines and to inform the development and implementation of strategies and technologies to improve air quality. The field is at an interesting juncture that may be framed with the question – “how low is low enough for air quality standards and guidance?”

Science should have an important role in addressing that question, however, scientific knowledge alone will not be sufficient to answer the question. Epidemiological studies with improved design and execution may give us some marginal improvement in estimates of the health risks associated with exposure to low ambient concentrations of various air constituents. It will be especially challenging to demonstrate the shape of the ambient concentration-response relationship at concentrations that include background levels. Rather than a singular focus on health effects attributable to air pollution as in the past, future epidemiological studies will better serve the needs of Society when

they consider more broadly the multiple factors that influence disease outcomes. A finding of relative risks of 1.1 or less for a specific air pollutant begs the question of what are the dominant risk factors driving the baseline health risks.

It has become commonplace for scientists using cellular and molecular approaches to argue for more research using their tools to investigate the mechanisms of action of various air pollutants. Based on the evidence that air pollution in the highest income countries of the world is not a major influence on the health of the population it is appropriate to ask whether a more detailed understanding of this specific risk factor, air pollutants, will be of value in guiding critical societal decisions. An alternative would be to focus on the broader range of risk factors that influence the baseline risk.

Irrespective of the specifics of the science available, it can be anticipated that increasing attention will focus on how science is used to decide on the level and statistical form of air quality standards and guidelines. The issue is relevant to both the setting of Standards and Guidelines, however, it is especially critical as it relates to standards which are legally binding. A review of all the existing Guidelines and Standards for criteria or common air pollutants suggests they have been based on judgment. In some cases, those judgments were made exclusively by scientists, that is, the WHO Air Quality Guidelines. In other cases, as in the setting of National Ambient Air Quality Standards by the US EPA, the decisions were made by appointed government officials and informed by the science reviewed by the Clean Air Scientific Advisory Committee. As best it can be determined none of the Guidelines or Standards were set at a specific defined level of risk. This is remarkable since each level (and its associated statistical form) for a Guideline or Standard reflects a level of acceptable risk: above this level and statistical form the health risks are *not* acceptable and below this level and statistical form the health risks of exposure are deemed acceptable. If scientists argue that there is a linear, no-threshold relationship between ambient concentrations of an air pollutant and associated health responses down to background concentrations, then no concentration is free of some level of calculated excess risk.

In this author's view, decisions on acceptable health risks for the public should be informed by the best

science available. However, scientists alone are not in a position to make broad policy decisions on acceptable risk. Indeed, pronouncements by individual scientists or scientific committees need to be recognized as a statement of the individual preference for a particular policy outcome or the collective view of the Committee. The current worldwide economic crisis results in every major financial investment made by either public or private entities being considered in the context of the future return on the investment in terms of the public or private good that will result. The focus increasingly is on choices that must be made between alternative uses of funds, irrespective of whether the funding is from the public or private sector. Clean air is universally viewed as a right. The question increasingly for policy makers, as representatives of the public, is how clean is clean enough? This question cannot be addressed in isolation; it should be addressed in the context of all the risk factors influencing human health. The question for scientists is how to create the scientific knowledge on air pollutants and other risk factors that affect human health to inform these important societal decisions.

## Bibliography

### Primary Literature

1. Adams WC (2002) Comparison of chamber and face mask 6.6-hour exposure to 0.08 ppm ozone via square-wave and triangular profiles on pulmonary responses. *Inhal Toxicol* 15:265–281
2. Adams WC (2003) Relation of pulmonary responses induced by 6.6 h exposures to 0.08 ppm ozone and 2-h exposures to 0.30 ppm via chamber and face-mask inhalation. *Inhal Toxicol* 15:745–759
3. Adams WC (2006) Comparison of chamber 6.6 h exposures to 0.04–0.08 ppm ozone via square-wave and triangular profiles on pulmonary responses. *Inhal Toxicol* 18:127–136
4. *American Trucking Associations, Inc. v. U.S. Environmental Protection Agency* (1999) DC circuit, 19 May 1999
5. Ayres JG, Harrison RM, Maynard R, McClellan RO, Nicols GL (2010) Environmental medicine in context, chapter 1. In: Ayres JG, Harrison RM, Maynard R, Nichols GL (eds) *Textbook of environmental medicine*. Hodder Education, London, pp 3–21
6. Ayres JG, Harrison RM, Maynard R, McClellan RO, Nicols GL (eds) (2010) *Textbook of environmental medicine*. Hodder Education, London
7. Bachmann J (2007) Will the circle be unbroken: a history of the U.S. National Ambient Air Quality Standards. *J Air Waste Manag Assoc* 57:652–697



8. Bell ML, McDermott A, Zeger SL, Samet JM, Dominici F (2004) Ozone and short-term mortality in 95 US urban communities, 1987–2000. *JAMA* 292(19):2372–2378
9. Bell ML, Dominici F, Samet JM (2005) A meta-analysis of time-series studies of ozone and mortality with comparison to the national morbidity, mortality, and air pollution study. *Epidemiology* 16(4):436–445
10. Berry MA (1984) A method for examining policy implementation: a study of decision-making for the National Ambient Air Quality Standards, 1964–1984. U.S. Environmental Protection Agency, Research Triangle Park, NC
11. Breyer S (1982) Regulation and reform. Harvard University Press, Cambridge
12. Breyer S (1993) Breaking the vicious circle: toward effective risk regulation. Harvard University Press, Cambridge, MA (Based on 1992 Oliver Wendell Holmes lectures at Harvard Law School)
13. Brown JS, Bateson TF, McDonnell WF (2008) Effects of exposure to 0.06 ppm ozone in FEV<sub>1</sub> in humans: a secondary analysis of existing data. *Environ Health Persp* 116:1023–1026
14. CAA (1963) The Clean Air Act of 1963
15. CAA (1970) The Clean Air Act Amendments of 1970-P.L. 91–04 (December 31, 1970), Plus Technical Amendments Made by P.L. 92–157 (November 18, 1971)
16. CAA (1977) The Clean Air Act Amendments of 1977
17. CAA (1990) The Clean Air Act Amendments to 1990, Public Law No. 101–549; 104 STAT 2399, 1990
18. Clinton WJ, President (1993) Executive Order 12866 – regulatory planning and review. *Fed Regis* 58:51735
19. Clinton WJ, President (1997) Memorandum for the Administrator of the Environmental Protection Agency Implementation of Revised Air Quality Standards for Ozone and Particulate Matter
20. Dockery DW, Pope CA III, Xu X et al (1993) An association between air pollution and mortality in six U.S. cities. *N Engl J Med* 329:1753–1759
21. Emanuel R (2009) Assistant to the President and Chief of Staff, Memorandum to the Heads of Executive Departments and Agencies. *Fed Regist* 74(13):4435–4436
22. Enstrom JE (2005) Fine particulate pollution and total mortality among elderly Californians, 1973–2002. *Inhal Toxicol* 17(14):803–816
23. Ferris BG Jr, Speizer FE, Spengler JD et al (1979) Effects of sulfur oxides and respirable particles on human health: methodology and demography of populations in study. *Am Res Respir Dis* 120:767–779
24. Follinsbee LJ, McDonnell WF, Horstman DH (1988) Pulmonary function and symptom responses after 6–6-hour exposure to 0.12 ppm ozone with moderate exercise. *JAPCA* 38:28–35
25. Haagen-Smit AJ (1952) Chemistry and physiology of Los Angeles smog. *Ind Eng Chem* 44:1342–1346
26. Horstman DH, Follinsbee LJ, Ives PJ, Abdul-Salaam S, McDonnell WF (1990) Ozone concentration and pulmonary response relationships for 6.6-hour exposures with five hours of moderate exercise to 0.08, 0.10, and 0.12 ppm. *Am Rev Respir Dis* 142:1158–1163
27. Kennedy D (2005) Risk versus risk. *Science* 309:2137
28. Krewski D, Burnett RT, Goldberg, Hoover K, Siemiatycki J, Jerrett M, Abrahamowicz M, White WH (2000) Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of particulate air pollution and mortality. Special report, Health Effects Institute, Cambridge, MA
29. Laden F, Schwartz J, Speizer FE, Dockery DW (2006) Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard six cities study. *Am J Respir Crit Care Med* 173:667–672
30. Library of Congress (1974) A Legislative History of the Clean Air Amendments of 1970, vol 1; Serial Number 93–18: Environmental Policy Decision, Congressional Research Service, Library of Congress, Washington, DC
31. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL (eds) (2006) Global burden of disease and risk factors. Oxford University Press, New York
32. Martin R, Symington L (1968) A guide to the Air Quality Act of 1967. *Law Contemp Probl* 33:239–245
33. Martineau R, Novello D (2004) The clean air act handbook, 2nd edn. American Bar Association, Chicago, Illinois
34. McClellan RO (1997) Use of mechanistic data in assessing human risks of exposure to particles. *Env Health Persp* 105(5):1363–1372
35. McClellan RO (1999) Human health risk assessment: a historical overview and alternative paths forward. *Inhal Toxicol* 11:477–518
36. McClellan RO (1999) Ambient airborne particulate matter: toxicology and standards. In: Gardner DE, McClellan RO, Crapo JD (eds) Toxicology of the lung, 3rd edn. Taylor and Francis, Washington, DC, p 289
37. McClellan RO (1999) Developing risk assessments for airborne materials. In: Gardner DE, McClellan RO, Crapo JD (eds) Toxicology of the lung, 3rd edn. Taylor and Francis, Washington, DC, p 599
38. McClellan RO (1999) Health risk assessments and regulatory considerations for air pollutants. In: Swift DL, Foster WM (eds) Air pollutants and the respiratory tract. Marcel Dekker, New York, pp 89–338
39. McClellan RO (2002) Setting ambient air quality standards for particulate matter. *Toxicology* 181–182:329–347
40. McClellan RO (2010) Hazard and risk: assessment and management, chapter 4. In: Ayres JG, Harrison RM, Maynard R, Nichols GL (eds) Textbook of environmental medicine. Hodder Education, London, pp 59–88
41. McClellan RO (2011) Role of science and judgment in setting National Ambient Air Quality Standards: how low is low enough? *Air Qual Atmos Health J* (Published on-line, June 1, 2011)
42. McClellan RO, Jackson TE (1999) Carcinogenic responses to air pollutants. In: Holgate ST, Samet JM, Koren HS, Maynard RL (eds) Air pollution and health. Academic, San Diego, pp 381–413

43. McClellan RO, Frampton MW, Koutrakis P, McDonnell WF, Moolgavkar S, North DW, Smith AE, Smith RL, Utell MJ (2009) Critical considerations in evaluating scientific evidence of health effects of ambient ozone: a conference report. *Inhal Toxicol* 21(Suppl 2):1–36
44. McDonnell WF, Kehri HR, Abdul-Salaam S et al (1991) Respiratory response of humans exposed to low levels of ozone for 6.6 hours. *Arch Environ Health* 46:145–150
45. McDonnell WF, Stewart PW, Smith MV (2007) The temporal dynamics of ozone-induced FEV<sub>1</sub> changes in humans: an exposure-response model. *Inhal Toxicol* 19:483–494
46. New York Times (1997) Clinton sharply tightens air pollution regulations despite concern over costs. *NY Times*, June 26, 1997
47. NRC (National Research Council) (2006) Health risks for exposure to low levels of ionizing radiation. BEIR VII – phase 2. Authored by Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation National Research Council
48. NRC (National Research Council) (2008) Estimating mortality risk reduction and economic benefits from controlling ozone air pollution. National Academies Press. [http://books.nap.edu/catalog.php?record\\_id=12198&utm\\_medium=email&utm\\_source=National%20Academies%20Press&utm\\_campaign=New+from+NAP+4.29.08-May-7-2008&utm\\_content=Downloader&utm\\_term=](http://books.nap.edu/catalog.php?record_id=12198&utm_medium=email&utm_source=National%20Academies%20Press&utm_campaign=New+from+NAP+4.29.08-May-7-2008&utm_content=Downloader&utm_term=). Accessed 2 Dec 2011
49. NRDC (1976) *Natural Resources Defense Council v. Train*, 411 F. Suppl 864 (S.D.N.Y.) aff'd 545 F. 2d 320 (2nd Cir., 1976)
50. Özkaynak H (1999) Exposure assessment. In: Halgate ST, Samet JM, Koren HS, Maynard RL (eds) *Air pollution and health*. Academic, London/San Diego, pp 149–162
51. Peacock M (2006) Memorandum from Marcus Peacock, Deputy Administrator to Dr. George Gray and Bill Wehrun. Process for reviewing the National Ambient Air Quality Standards. [http://www.epa.gov/ttnnaaqs/pdfs/memo\\_process\\_for\\_reviewing\\_naaqs.pdf](http://www.epa.gov/ttnnaaqs/pdfs/memo_process_for_reviewing_naaqs.pdf) (accepted 2007). Accessed 7 Dec 2006
52. Pedersen WF (1975) Formal records and informal rulemaking. *Yale Law J* 38:38–88
53. Pope CA III, Dockery DW (1999) Epidemiology of particle effects. In: Holgate ST, Samet JM, Loren HS, Maynard RL (eds) *Air pollution and health*. Academic, London, pp 673–705
54. Pope CA III, Burnett RJ, Thun MJ, Calle EE, Krewski D, Ito K, Thurston CD (2002) Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA* 287(9):1132–1141
55. Pruss-Ustun A, Corvalan C (2006) Towards estimation of the environmental burden of disease. World Health Organization, Copenhagen, Denmark
56. Rhomberg LR, Goodman JE, Haber LT, Dougson M, Andersen MT, Klaunig JE, Meek B, Price PS, McClellan CSM (2011) Linear low-dose extrapolation for non-cancer health effects is the exception, not the rule. *Crit Rev Toxicol* 41(1):1–19
57. Schrenk HH, Heimann H, Clayton GD, Gafafer WM, Wexler H (1948) Air pollution in Donora, PA. *Epidemiology of the unusual smog episode of October 1948*. Public Health Bull 306; U.S. Government Printing Office, Washington, DC, 1949
58. Smith RL, Baowei X, Switzer P (2009) Reassessing the relationship between ozone and short-term mortality in U.S. urban communities. *Inhal Toxicol* 29(S2):37–61
59. Steenland K, Hu S, Walker J (2004) All-cause and cause-specific mortality by socioeconomic status among employed persons in 27 U.S. States, 1984–1997. *Am J Public Health* 94: 1037–1042
60. Stern AC (1982) History of air pollution legislation in the United States. *J Air Poll Control Assoc* 32:44–61
61. Sunstein CR (2004) Lives, life years, and willingness to pay. *Columbia Law Rev* 104(1):205–252
62. Sunstein CR (2011) Letter from Cass Sunstein, Administrator, Office of Information and Regulatory Affairs, Office of Management and Budget to Administrator Lisa Jackson, U.S. Environmental Protection Agency on “Reconsideration of the 2008 Ozone Primary and Secondary National Ambient Air Quality Standard” dated September 2, 2011
63. Taube G (1995) Epidemiology faces its limits. *Science* 269: 164–169
64. U.S. Environmental Protection Agency (1971) National primary and secondary Ambient Air Quality Standards. *Fed Regist* 36:8186
65. U.S. Environmental Protection Agency (1997) National Ambient Air Quality Standards for ozone. *Fed Regist* 62(138): 38856–38896
66. U.S. Environmental Protection Agency (1997) National Ambient Air Quality Standards for particulate matter: final rule. *Fed Regist* 62(138):38652–38760
67. U.S. Environmental Protection Agency (1997) EPA’s Regulatory Impact Analyses (RIA) for the 1997 ozone and PM NAAQS and proposed regional haze rule. <http://www.epa.gov/ttn/oarpg/naaqsfin/ria.html>. Accessed 2 Dec 2011
68. U.S. Environmental Protection Agency (2004) Air quality criteria for particulate matter. Final report, Oct 2004. <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=87903>. Accessed 2 Dec 2011
69. U.S. Environmental Protection Agency (2005) Particulate matter health risk assessment for selected urban areas. U.S. Environmental Protection Agency, Research Triangle Park, NC, EPA 452/R-05-007A
70. U.S. Environmental Protection Agency (2005) Review of the National Ambient Air Quality Standards for particulate matter: policy assessment of scientific and technical information, OAQPS Staff Paper. U.S. Environmental Protection Agency, Research Triangle Park, NC, EPA-452/R-05-005a
71. U.S. Environmental Protection Agency (2006) Air quality criteria for ozone and related photochemical oxidants (2006 Final). EPA/600/R-05-004aF-cF. <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=149923>. Accessed 2 Dec 2011
72. U.S. Environmental Protection Agency (2006) Air quality criteria for lead. Final report. <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=158823>. Accessed 2 Dec 2011

73. U.S. Environmental Protection Agency (2006) National Ambient Air Quality Standards for lead: final rule. 40 CFR Parts 50, 51, 53 and 58. Fed Regist 73:66964–67062
74. U.S. Environmental Protection Agency (2006) National Ambient Air Quality Standards for particulate matter: proposed rule. Fed Regist 71:2620–2708
75. U.S. Environmental Protection Agency (2006) National Ambient Air Quality Standards for particulate matter: final rule. Fed Regist 71:61144–61233
76. U.S. Environmental Protection Agency (2006) Regulatory impact analysis for the National Ambient Air Quality Standard for particulate matter (Issues October 6, 2006). [http://www.epa.gov/ttn/naaqs/standards/pm/s\\_pm\\_cr\\_fr.html](http://www.epa.gov/ttn/naaqs/standards/pm/s_pm_cr_fr.html). Accessed 2 Dec 2011
77. U.S. Environmental Protection Agency (2007) Ozone health risk assessment for selected urban areas. U.S. Environmental Protection Agency, Research Triangle Park, NC, EPA 452-R-07-009
78. U.S. Environmental Protection Agency (2007) Ozone population exposure analysis for selected urban areas. U.S. Environmental Protection Agency, Research Triangle Park, NC, EPA 452-R-07-010
79. U.S. Environmental Protection Agency (2007) Review of the National Ambient Air Quality Standards for ozone: policy assessment of scientific and technical information
80. U.S. Environmental Protection Agency (2007) National Ambient Air Quality Standards for ozone: proposed rule 40 CFR part 50. Fed Regist 72:37818–37919
81. U.S. Environmental Protection Agency (2008) National Ambient Air Quality Standards for ozone: final rule 40 CFR parts 50 and 58. Fed Regist 73:16436–16514
82. U.S. Environmental Protection Agency (2008) Final ozone NAAQS regulatory impact analysis, EPA-452/R-08-003. [http://www.epa.gov/ttnecas1/regdata/RIAs/452\\_R\\_08\\_003.pdf](http://www.epa.gov/ttnecas1/regdata/RIAs/452_R_08_003.pdf). Accessed 2 Dec 2011
83. U.S. Environmental Protection Agency (2009) Risk and exposure assessment to support the review of the SO<sub>2</sub> primary National Ambient Air Quality Standards. U.S. Environmental Protection Agency, Research Triangle Park, NC, EPA/452/R-09-007
84. U.S. Environmental Protection Agency (2009) Integrated science assessment for particulate matter (Final report). <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=216546>. Accessed 2 Dec 2011
85. U.S. Environmental Protection Agency (2010) Proposed revisions to National Ambient Air Quality Standards for ozone: proposed release. Fed Regist 73:2938–3052
86. U.S. Environmental Protection Agency (2010) Primary National Ambient Air Quality Standards for nitrogen dioxides: final rule. Fed Regist 75:6474
87. U.S. Environmental Protection Agency (2010) Primary National Ambient Air Quality Standard for sulfur dioxide: final rule. Fed Regist 75:35520–35603
88. U.S. Environmental Protection Agency (2011) Integrated science assessment for ozone and related photochemical oxidants (first external review draft). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-10/076A
89. U.S. Environmental Protection Agency (2011) Review of National Ambient Air Quality Standards for carbon monoxide: final rule. Fed Regist 76:54204–54343
90. Wang H, Jacob DJ, Le Sager P, Streets DG, Park RJ, Gilliland AB, van Donkelaar A (2009) Surface ozone background in the US: Canadian and Mexican pollution influences. Atmos Environ 43:1310–1319
91. White House Press Release (2011) Statement by the President on the Ozone National Ambient Air Quality Standards
92. White RH, Cote I, Zeise L, Fox M, Dominici F, Burke TA, White PD, Hattis DB, Samet JM (2009) State-of-the-science workshop report: issues and approaches in low dose–response extrapolation for environmental health risk assessment. Environ Health Perspect 117(2). doi:10.1289/ehp.11502. <http://www.ehponline.org/members/2008/1150211502.pdf>
93. *Whitman v. American Trucking Associations* (2001) 531 U.S. 457, 121 S. Ct. 903149L. Ed. 2d1
94. WHO (World Health Organization) (1987) Air quality guidelines for Europe. World Health Organization Regional Office for Europe, Copenhagen (WHO Regional Publications, European Series No. 23). <http://www.who.int/mediacentre/factsheets/fs313/en/index.html>. Accessed 2 Dec 2011
95. WHO (World Health Organization) (2000) Air quality guidelines for Europe, 2nd edn. WHO Regional Publications, European Series No. 91. [http://www.euro.who.int/\\_\\_data/assets/pdf\\_file/0005/74732/E71922.pdf](http://www.euro.who.int/__data/assets/pdf_file/0005/74732/E71922.pdf). Accessed 2 Dec 2011
96. WHO (World Health Organization) (2005) WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide, global update 2005. Summary of risk assessment. [http://www.who.int/phe/health\\_topics/outdoorair\\_aqq/en/](http://www.who.int/phe/health_topics/outdoorair_aqq/en/). Accessed 2 Dec 2011
97. Wolff GT (1996) Letter to Carol M. Browner, U.S. Environmental Protection Agency Administrator. Closure on Draft Office of Air Quality Planning and Standards (OAPQS Staff Paper). Review of the National Ambient Air Quality Standards for particulate matter: policy assessment of science and technical information

## Books and Reviews

- Donaldson K, Borm P (eds) (2007) Particle toxicology. CRC Press/Taylor and Francis, New York
- Finlayson-Pitts BJ, Pitts JN Jr (2000) Chemistry of the upper and lower atmosphere: theory, experiments, and applications. Academic, San Diego, California
- Gardner DE (2005) Toxicology of the lung, 4th edn. CRC Press, Boca Raton, Florida
- Gardner DE, Crapo JD, McClellan RO (eds) (1999) Toxicology of the lung, 3rd edn. Taylor and Francis, Washington, DC
- Gehr P, Heyder J (eds) (2000) Particle-lung interactions. Marcel Dekker, New York
- Godish T (2003) Air quality, 4th edn. Taylor and Francis, Philadelphia, PA

- Harrison RM, Hester RE, Harrison RM (eds) (2009) Air quality in urban environments (Issues in environmental science and technology). Royal Society of Chemistry, London, UK
- Hayward SF, Schwartz JM (eds) (2007) Air quality in America. AEI Press, Washington, DC
- Hodgson E (2010) A textbook of modern toxicology. Wiley, Hoboken, New Jersey
- Holgate ST, Samet JM, Koren HS, Maynard RL (eds) (1999) Air pollution and health. Academic, San Diego
- Klaassen CD (ed) (2007) Casarete and Doull's toxicology, 7th edn. McGraw-Hill, New York, NY
- Lippmann M (ed) (2009) Environmental toxicants, human exposures and their health effects, 3rd edn. Wiley, Hoboken
- McClellan RO, Henderson RF (eds) (1995) Concepts in inhalation toxicology, 2nd edn. Taylor and Francis, Philadelphia, Pennsylvania
- McMurry PH, Shepherd MF, Vickery JS (eds) (2004) Particulate matter science for policy makers – a NARSTO assessment. Cambridge University Press, Cambridge, MA
- Morris JB, Shusterman DJ (eds) (2010) Toxicology of the nose and upper airways. Informa Health Care, London
- Phalen RF (2002) The particulate air pollution controversy: a case study and lessons learned. Kluwer Academic, Boston, Massachusetts
- Phalen RF, Phalen RN (2011) Introduction to air pollution science. Jones and Bartlett Learning, Burlington, MA
- Preining O, Davis EJ (eds) (2000) History of aerosol science: proceedings of the symposium on the history of aerosol science, Vienna, 31 Aug–2 Sept 1999. Verlag Der Österreichischen Akademie Der Wissenschaften, Wien
- Ruzer LS, Haley NH (eds) (2005) Aerosols handbook: measurement, dosimetry and health effects. CRC Press, Boca Raton, Florida
- Schäfer S, Marquardt H, Welsch F, McClellan R (1999) Toxicology. Academic, San Diego, California
- Seinfeld JH, Pandis SN (2000) Atmospheric chemistry and physics: from air pollution to climate change, 2nd edn. Wiley, New York, NY

## Air Quality, Surface Transportation Impacts on

PAUL BOULTER  
PAEHolmes, Sydney, Australia

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Historical Trends  
Air Pollution Pathways

Emission Processes  
Modelling Transport Emissions  
Pollutant Dispersion and Transformation  
Modelling Air Pollution from Transport  
Reducing Air Pollution from Transport  
Future Directions  
Bibliography

### Glossary

**DPF** Diesel particulate filter, the preferred and most effective technology for removing particulate matter from diesel exhaust. DPFs can be broadly divided into two types: “full flow” and “partial flow.” The latter are more permeable and are less effective, but need no maintenance. DPFs physically capture particles and can fill up rapidly. They must therefore be periodically “regenerated,” and several passive and active systems have been developed for this purpose.

**Emission standard** A value that specifies the maximum amount of a pollutant that can be released into the environment per unit activity (e.g., g/km or g/kWh).

**NO<sub>2</sub>** Nitrogen dioxide, a gaseous air pollutant produced as a result of combustion. NO<sub>2</sub> is also a precursor for a number of harmful secondary air pollutants such as ozone, and plays a role in the formation of acid rain. Long-term exposure to high concentrations of NO<sub>2</sub> causes adverse health effects, and levels regularly exceed air quality standards.

**O<sub>3</sub>** Ground-level ozone, a secondary pollutant produced by reactions between NO<sub>2</sub>, hydrocarbons, and sunlight. High levels of ozone are generally observed during hot, sunny weather, and concentrations tend to be higher in rural areas than in urban areas.

**PM<sub>10</sub>** Mass concentration of particles with an aerodynamic diameter of less than 10 µm. Particles smaller than 10 µm (also known as “thoracic” particles) can settle in the bronchi and lungs.

**PM<sub>2.5</sub>** Mass concentration of particles with an aerodynamic diameter of less than 2.5 µm, sometimes referred to as “fine” or “respirable” particles. Particles in this size range can penetrate deep into the lungs.

**Primary particles** Particles emitted directly to the atmosphere, typically as a result of combustion or abrasion processes.

**SCR** Selective catalytic reduction, an exhaust aftertreatment method that is used to reduce emissions of nitrogen oxides. It will be the main technology for enabling compliance with future emission standards.

**Secondary particles** Particles formed within the atmosphere from gas-phase precursors. These include particles originating from atmospheric oxidation of sulfur and nitrogen oxides, and their reaction products with ammonia, and from the oxidation of organic compounds.

**Ultrafine particles** Particles with a diameter of less than 0.1  $\mu\text{m}$ . These can pass through cell membranes and migrate from the lungs into other organs. There is evidence to indicate that they may be more damaging to the cardiovascular system than  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ .

## Definition of the Subject

Until the middle of the twentieth century poor air quality in populated areas was usually associated with domestic and industrial combustion, with coal burning leading to high concentrations of smoke and sulfur dioxide. A recent retrospective analysis has indicated that the most famous example of this type of pollution – the great London smog of 1952 – was responsible for 12,000 deaths [1]. However, during the 1940s and 1950s an altogether different type of pollution was observed in the United States. Haagen-Smit identified a mechanism by which air pollutants reacted with the Los Angeles sunlight to produce a noxious photochemical smog [2], and it became clear that the pollution originated not from the use of coal but from road vehicles.

In the subsequent decades urban air pollution has become synonymous with road transport, the link being attributable primarily to the growth in the ownership and use of private cars. Photochemical smog has remained a problem in some of the largest cities in the world, and many people are exposed to pollutant concentrations which exceed health-based air quality standards. It has also become evident that some transport-derived pollutants can travel thousands of kilometers before deposition and damage occur and, in the case of greenhouse gases, can have a global effect over long timescales. If transport systems are to become

environmentally sustainable, then the scale of such adverse effects will have to be greatly reduced.

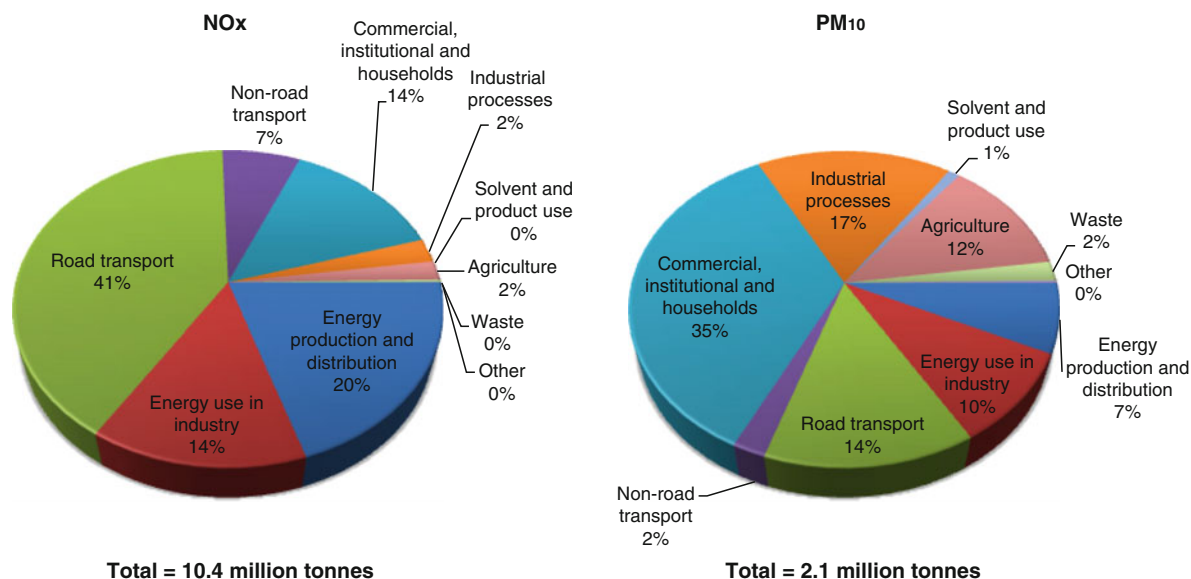
The aim of this entry is to provide an overview of the impacts of surface transport (road, rail, and shipping) on emissions and air quality. Different regions of the world have different air pollution problems, and not all countries and issues can be covered here. The entry refers mainly, but not exclusively, to the European Union (EU), and the emphasis is on local air pollution in populated areas. Inevitably, the entry deals predominantly with road transport, but some consideration is also given to the other modes.

## Introduction

Surface transport is an important, and in many respects, the most important source of air pollution. The pollutants released from road, rail, and waterborne traffic are implicated in a variety of direct and indirect detrimental impacts on amenity, health, ecosystems, and cultural heritage over different time periods and on different geographical scales. The effects of transport pollution can be very local and immediate, such as the nuisance caused by a plume of black smoke from the exhaust pipe of a car or a diesel train. More seriously, repeated exposure to exhaust gases and particles is linked to aggravated respiratory and cardiovascular disease, changes to lung tissue, changes in the function of the nervous system, and premature mortality [3, 4]. Threats to cultural heritage posed by atmospheric pollutants include the soiling of historical buildings – a common sight alongside busy urban roads – and indoor damage to works of art and museum exhibits. Transport emissions also contribute to the regional degradation of air quality, to eutrophication, and to acidification (see entry on ► [Regional Air Quality](#)). Moreover, stable pollutants, including the greenhouse gases carbon dioxide ( $\text{CO}_2$ ), methane, and nitrous oxide can contribute to environmental problems on a global scale for decades and even centuries.

The pollutants that are currently causing the greatest concern in terms of air quality, primarily because of their impact on human health, are airborne particulate matter (PM), nitrogen dioxide ( $\text{NO}_2$ ), and ground-level ozone ( $\text{O}_3$ ). Surface transport is an important contributor to all three [5], but road





**Air Quality, Surface Transportation Impacts on. Figure 1**

NO<sub>x</sub> and PM<sub>10</sub> emissions in 2008 by sector, based on data submitted by EU-27 countries to CLRTAP [8]

transport is the main culprit, especially in population centers, and it presents some of the greatest challenges in terms of sustainability.

The importance of road transport as a source of pollution can be illustrated by reference to sectoral emissions in Europe. Figure 1 shows emission data for the EU-27 countries, based on submissions to the UNECE Convention on Long-range Transboundary Air Pollution (CLRTAP). In 2008, road transport was the largest contributor to NO<sub>x</sub> emissions in Europe (41%), and was also a major contributor to PM<sub>10</sub> emissions. By comparison, non-road transport was responsible for just 7% of total NO<sub>x</sub> emissions and only 2% of total PM<sub>10</sub> emissions. However, it is important to observe that the totals in Fig. 1 are based on national reporting, and do not include international aviation or international maritime shipping. The latter is an important source of NO<sub>x</sub> and sulfur oxides (SO<sub>x</sub>). Recent studies have indicated that around 15% of global NO<sub>x</sub> emissions and 5–8% of global SO<sub>x</sub> emissions are attributable to oceangoing ships [6, 7].

At the local level, the impacts of road transport are exacerbated by the proximity of vehicles (and their emissions) to the population, although the other surface transport modes can contribute

significantly to poor air quality at specific sites such as busy railway stations, underground systems, and large ports.

Rail transport only contributes a small proportion to total emissions from the transport sector, but individual diesel locomotives have large engines, and exhaust aftertreatment equipment is not currently used on traction units. Elevated pollutant concentrations have therefore been observed where diesel locomotives are running with the engine at idle in enclosed stations [9]. Several studies have addressed the exposure of passengers to air pollution in underground rail systems, where pollutant concentrations appear to be considerably higher than those above ground. For example, consistently high concentrations of PM<sub>10</sub> and PM<sub>2.5</sub> have been reported [10–12].

Many major ports are also major cities, and so emissions from seagoing ships in such ports can directly affect a large population [13]. Ships at berth are the main source of shipping emissions in ports; the time spent at berth is typically 1 or more days, whereas the time spent maneuvering is typically only a few hours [14]. Moreover, almost 70% of maritime ship emissions occur within 400 km of land, and ships therefore have the potential to contribute significantly to pollution in coastal communities [13].

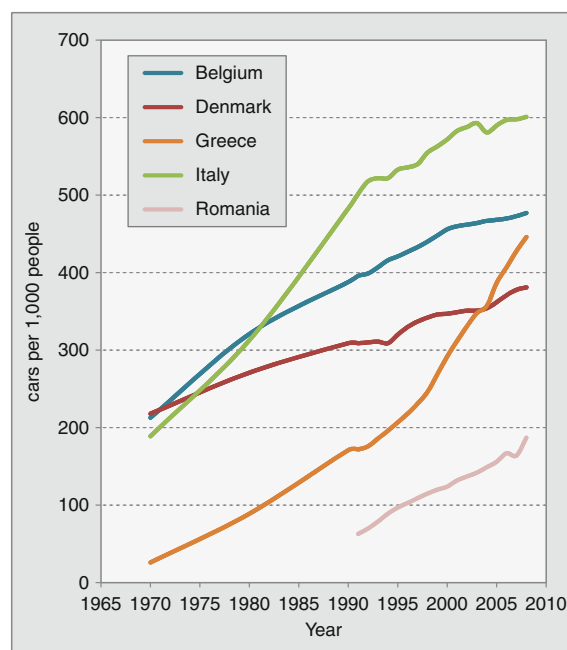
On the global scale, PM emissions from marine shipping are thought to be responsible for around 60,000 deaths annually, with impacts concentrated in coastal regions on major trade routes. Most mortality effects are seen in Asia and Europe where high populations and high shipping-related PM concentrations coincide [13].

The literature on transport-related air pollution is very extensive, and it is not possible to cover every aspect in depth in this entry. The emphasis is therefore on key areas such as historical trends in transport activity and emissions, the processes of pollutant formation, and modelling approaches. Importantly, given the context of this encyclopedia, a substantial part of the entry is devoted to policies and measures which can be implemented to reduce the impacts of surface transport on air quality, and thus contribute to sustainability.

### Historical Trends

The general trend in motorized transport has been one of unrelenting growth, driven primarily by the association of personal mobility with quality of life [15] and its consequence in terms of passenger car ownership. In the United States the number of cars grew from 140,000 in 1907 to more than 20 million in 1927 [15]. A similar pattern occurred in Europe several decades later – Fig. 2 shows car ownership in some European countries since 1970 – and is being repeated again now in parts of Asia.

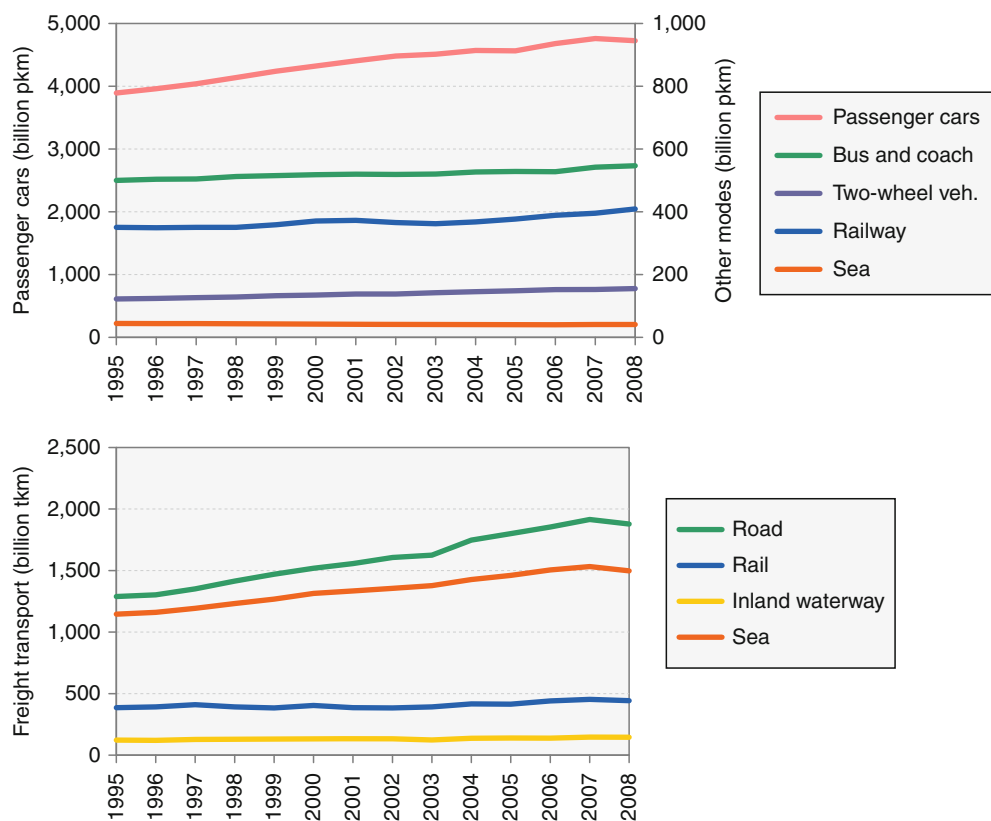
The recent evolution of transport activity in the EU-27 countries is shown in Fig. 3. The top graph shows passenger transport by mode in passenger-kilometers (pkm) (note the different scales), and the bottom graph shows freight transport in tonne-kilometers (tkm). Between 1995 and 2008, passenger transport by car increased by 21%. The growth in passenger transport was lower for buses/coaches (9%) and rail (17%), and there was a net reduction of 8% in passenger transport by sea. Although it is not of direct relevance to this entry, it should be recognized that during the same period passenger transport by air increased by 62% (see entry on ► [Aviation and Atmosphere](#)). Road freight in the EU-27 increased by almost 50% between 1995 and 2008, and there was also substantial growth in the transport of freight by rail (15%), inland waterways (19%) and shipping (30%).



**Air Quality, Surface Transportation Impacts on.** Figure 2 Passenger car ownership in selected European countries (Adapted from [16, 17])

In response to similar trends in all industrialized countries and their implications for air quality, legislation and strategies have been put in place to combat transport-related emissions. Calculations have established that emissions from road transport in the EU peaked in the early 1990s [18] but have been reducing since then as controls on fuel and emissions have tightened (Fig. 4). Emissions of regulated pollutants are likely to continue to decrease in the future as new vehicles, locomotives, and vessels, which comply with more stringent emission limits, replace older ones, although the reductions will be offset to some extent by growth in transport activity.

There have been some notable successes. In spite of the continued growth in traffic in Europe, the emissions and atmospheric concentrations of some pollutants – such as lead, carbon monoxide and benzene – have decreased so dramatically that they are no longer considered to be a major threat to human health. However, in many urban areas the concentrations of NO<sub>2</sub> and PM<sub>10</sub> still frequently exceed health-based limit values. According to the European Environment Agency, in 2004, more than 20% of the European



**Air Quality, Surface Transportation Impacts on. Figure 3**

Passenger and freight transport in the EU-27 countries by mode between 1995 and 2008 (Adapted from [17])

urban population were exposed to ambient  $\text{NO}_2$  concentrations above the annual mean limit value ( $40 \mu\text{g}/\text{m}^3$ ) [19]. Furthermore,  $\text{NO}_2$  and  $\text{PM}_{10}$  concentrations at many roadside monitoring sites are not decreasing [19–21]. The problem is not confined to Europe; for example, a similar problem has been observed in Japan [22]. Some possible reasons for this are explored later in this entry. A further concern is that emissions of non-exhaust particles – especially those arising from road vehicle tire and brake wear – are not subject to legislation or control, and are currently increasing in line with traffic growth (bottom graph in Fig. 4).

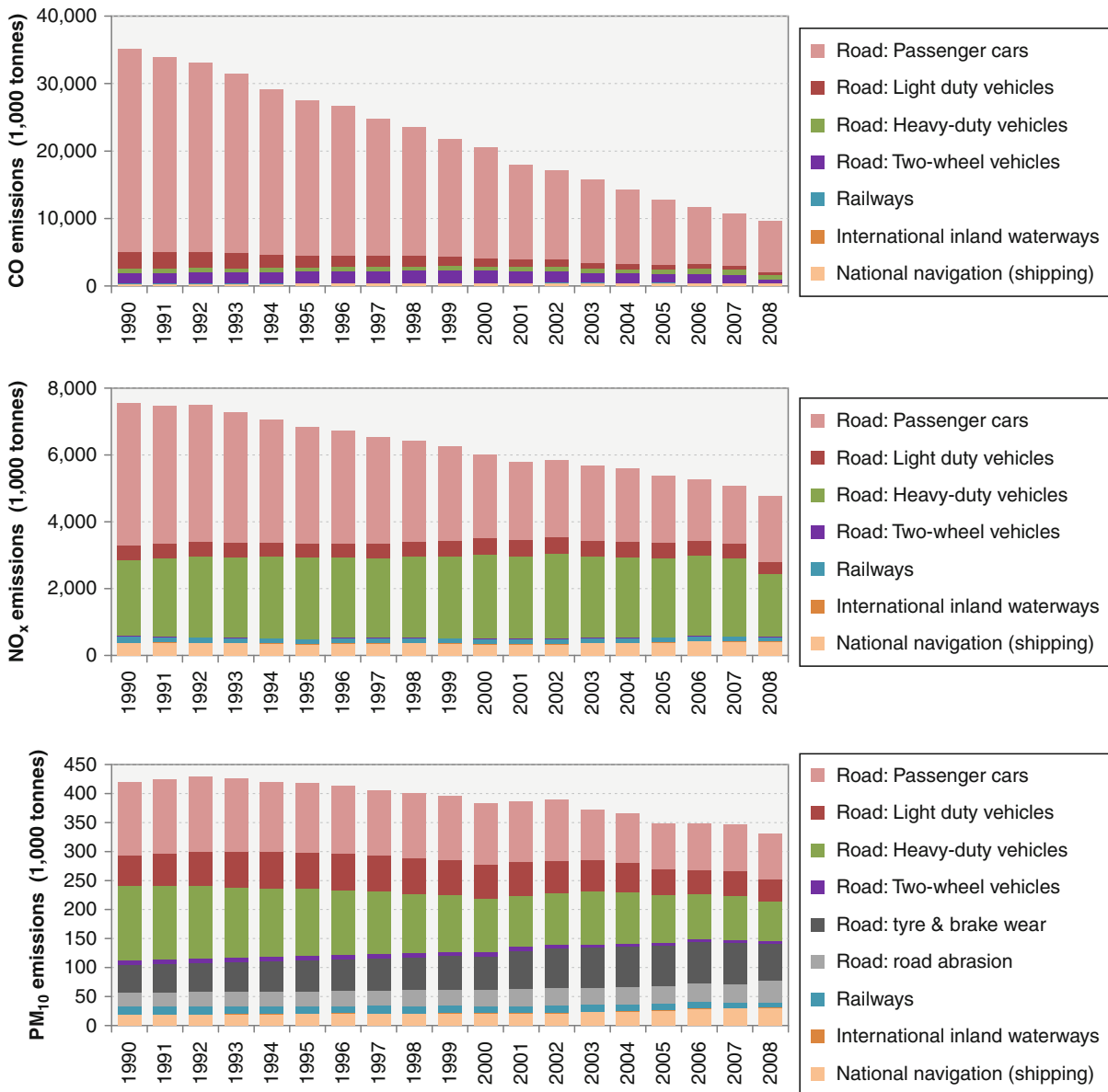
### Air Pollution Pathways

The links between transport, emissions, air pollution, and adverse impacts are not straightforward. To illustrate this, the complex pathway from emissions to one specific type of impact – the health

outcomes of road transport pollution – is summarized in Fig. 5.

Many different pollutants are emitted by road traffic, and the strength of the source depends on the emission factors for the vehicles and the characteristics of the traffic (e.g., volume, speed and composition). The emitted (primary) pollutants are then dispersed in the ambient air according to the local topography and meteorology. The main direct impacts of primary pollutants are near to the point of emission; further away, concentration decreases rapidly as a result of dispersion and dilution. However, many of the primary pollutants react together, and with pollutants from other sources, to form secondary pollutants. Because of the time required for their formation, the concentrations of secondary pollutants are not always highest near the emission source.

The overall exposure of individuals to air pollutants is dependent upon the types of activity in which they

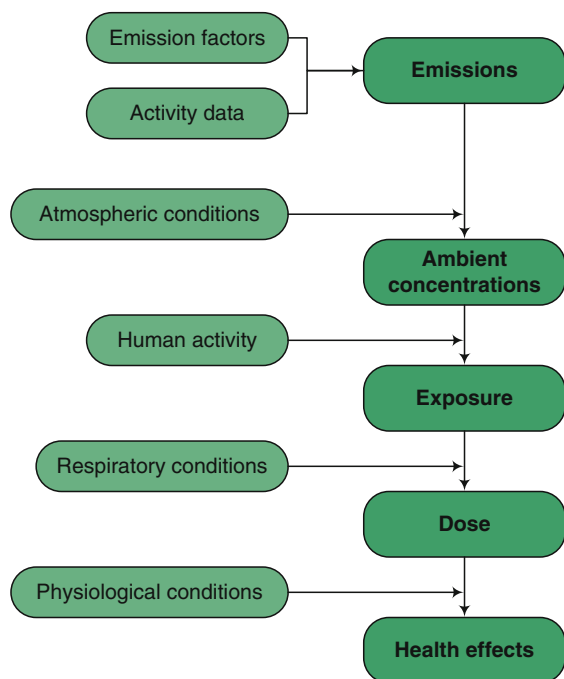


**Air Quality, Surface Transportation Impacts on. Figure 4**

CO, NO<sub>x</sub>, and PM<sub>10</sub> emissions by transport mode (excluding aviation) between 1990 and 2008, submitted by EU-27 countries to CLRTAP [8]

are engaged, the locations of those activities, and the pollutant concentrations in those locations. One way of determining “integrated” personal exposure is through the use of a “microenvironment” model [24]. The term microenvironment describes a location with a distinct level of pollution, such as outdoors in the street, indoors at home, in transit, at the workplace, etc. The

fraction of time spent in each microenvironment allows the calculation of integrated exposure. However, the calculation of such an integral is often not possible because the pollutant concentrations in the different microenvironments are not known. The term “average exposure” is therefore commonly used, and this is normally taken to mean the pollutant concentration



**Air Quality, Surface Transportation Impacts on.** Figure 5 Simplified pathway from transport emission to health effect (Adapted from [23])

over a specified period (e.g., annual mean) at an outdoor location which is broadly representative of where people are likely to spend time. This approach is used in the regulation of ambient air quality.

Once the pollutant has crossed a physical boundary (e.g., alveolar epithelial cells), the concept of “dose” is used [25]. The dose is the mass of material absorbed or deposited in the body during an interval of time, and depends on the respiratory activity of the individuals concerned. Responses to doses – the actual health effects – can also vary from person to person, depending on physiological conditions.

## Emission Processes

### Combustion

Most vehicles are powered by internal combustion engines (ICEs). Notable exceptions include vehicles powered by electricity drawn from an external grid (as in the case of rail locomotives), from on-board batteries (as in the case of road vehicles and waterborne vessels) or, more recently, from in situ generation involving fuel cells.

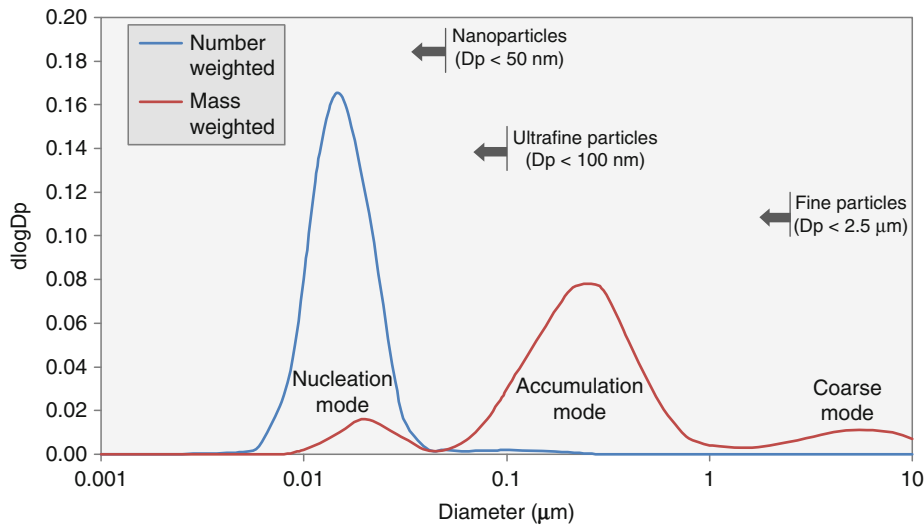
The principles of ICEs apply equally to road vehicles, rail locomotives, and waterborne vessels, the main differences being the size of the engine and the type of fuel used. In an ICE energy is derived from the burning of hydrocarbon fuel in air, with the main products being  $\text{CO}_2$  and water vapor. Incomplete combustion results in the production of CO, PM (especially in the case of diesel engines), and many different organic compounds. In addition, at the high temperatures and pressures found in the combustion chamber, some of the nitrogen in the air is oxidized, forming mainly nitric oxide (NO) with some  $\text{NO}_2$ . By convention, the sum total of oxides of nitrogen (i.e.,  $\text{NO} + \text{NO}_2$ ) is termed  $\text{NO}_x$ .

$\text{NO}$  formation proceeds via two main mechanisms: the “thermal” (Zel’dovich) and “prompt” mechanisms [26], with the former being responsible for more than 90% of emissions [27].  $\text{NO}_2$  is predominantly a secondary pollutant, its major source being the atmospheric oxidation of combustion-derived NO. The  $\text{NO}_2$  emitted directly from vehicles is referred to as “primary  $\text{NO}_2$ ,” and its importance will be described later.

Particles in diesel exhaust cover a range of sizes, and the shape of the size distribution depends on whether the weighting is by number or mass (Fig. 6). There are three distinct size modes: the nucleation mode (sometimes referred to as nuclei or nanoparticles), the accumulation mode, and the coarse mode. The nucleation mode has traditionally been defined as particles with a diameter of less than 50 nm. Accumulation mode particles range in size from around 50 nm to around 1  $\mu\text{m}$ , with particles smaller than 0.1  $\mu\text{m}$  being referred to as ultrafine particles. When a nucleation mode is observed (this is not always the case in engine exhaust), it contains many more particles than the accumulation mode, although because each particle is so small the total mass will not increase substantially. The coarse mode consists of particles larger than around 1  $\mu\text{m}$ .

The processes of particle formation during diesel combustion have been described in detail [26]. Carbonaceous spherules (soot) are initially created in the cylinder. A phase of particle growth then follows, involving the adsorption of gas-phase components, as well as coagulation and agglomeration. Almost all the particles found in the exhaust prior to dilution are





**Air Quality, Surface Transportation Impacts on. Figure 6**

Typical particle size distributions in vehicle exhaust. The y-axis is a normalized log scale (Adapted from [28])

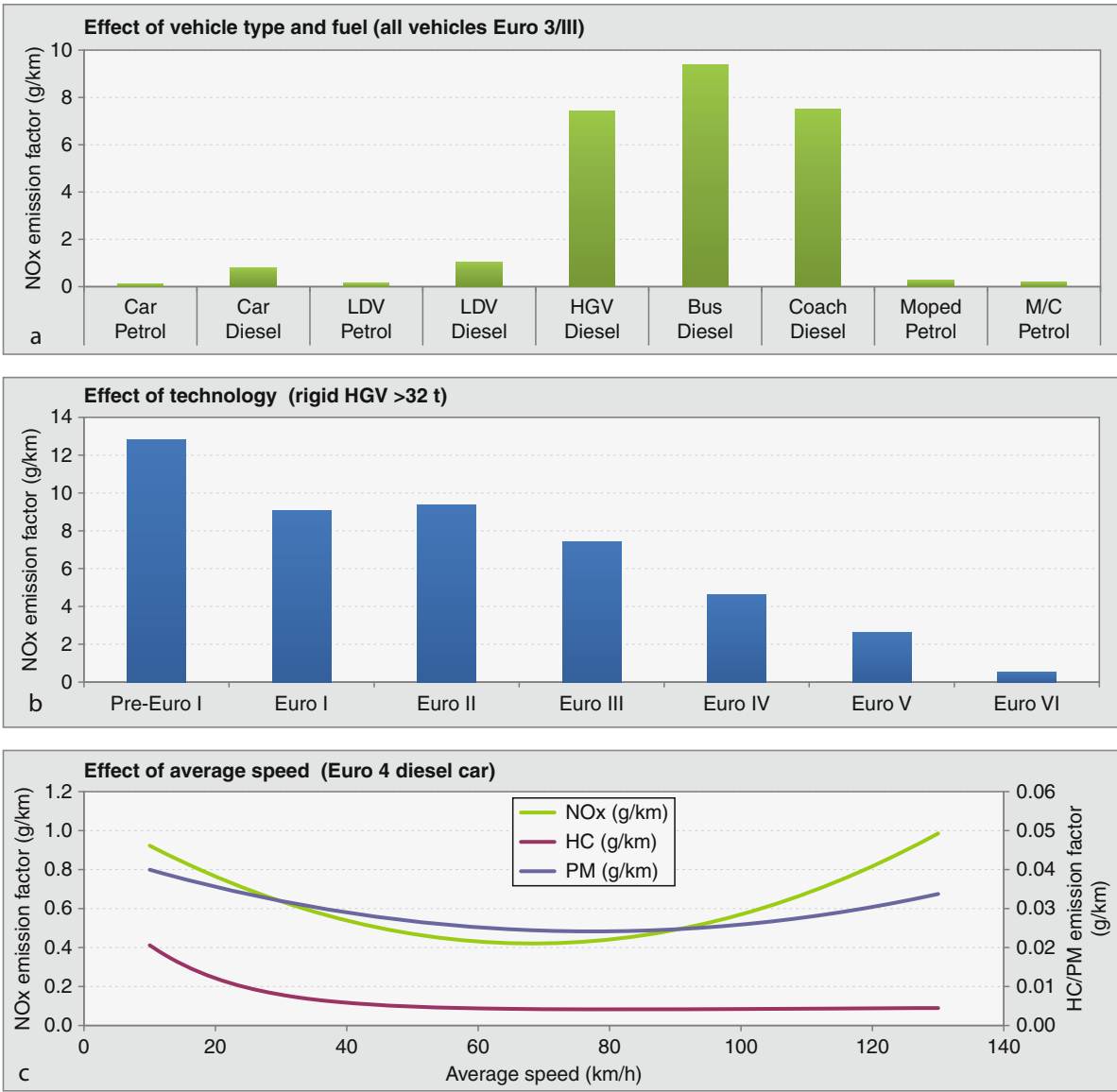
present as these carbonaceous agglomerates (hence, accumulation mode), with a small amount of metallic ash. The processes leading to the presence of nucleation mode particles in the exhaust during measurement are quite complex. Most nucleation mode particles are thought to originate from the condensation of volatile material (hydrocarbons, hydrated sulfuric acid and salts) in the exhaust gas during dilution, rather than during combustion itself, and their formation is a function of measurement parameters such as temperature, dilution ratio, residence time, and humidity [28–30], as well as fuel sulfur content [31, 32]. Particles in the coarse mode are formed by re-entrainment of material previously deposited on engine cylinder and exhaust system surfaces. Some of the pollutants emitted by fuel combustion (sulfur oxides and nitrogen oxides) are also precursors of secondary particles in the atmosphere.

The main factors that govern exhaust emissions are the vehicle type, the fuel used, the vehicle technology, and the driving conditions. The term “vehicle technology” can be used to refer to either a specific type of engine or exhaust aftertreatment, or more generally compliance with a particular emission standard. Some examples of these effects are shown for road vehicles in Fig. 7. Other factors include vehicle weight, road gradient, vehicle load, and the use of auxiliary equipment

such as air-conditioning. Another important consideration is the so-called dynamics of driving (see “[Modeling transport emissions](#)”).

The thermal state of the engine and exhaust emission-control system also has a strong influence on emissions from road vehicles. Most emissions occur under what are known as thermally stable or “hot” engine and exhaust conditions. The temperature of engine coolant during normal operation is typically between around 70°C and 90°C, whereas the temperature of the exhaust system reaches several hundred degrees centigrade. During the warm-up period following an engine start, emissions and fuel consumption are elevated as a result of incomplete combustion, the low conversion rate of pollutants in the catalyst, and increased viscous friction due to the low lubricant temperature in the engine and transmission. These “cold-start” emissions can constitute a significant proportion of total road transport emissions, particularly in urban areas. Cold-start emissions and fuel consumption are higher for lower engine starting temperatures which will, in turn, depend upon factors such as the length of time the vehicle has been parked and the prevailing ambient temperature.

Rail locomotives are mainly powered by either diesel or electricity. The nature of the pollution emitted from rail transport is dependent upon the type of



**Air Quality, Surface Transportation Impacts on. Figure 7**  
Factors influencing emissions: (a) vehicle and fuel type, (b) technology, and (c) average speed (Adapted from [33])

traction. In the case of diesel locomotives, pollutants are emitted directly into the local environment as result of combustion, whereas for electric locomotives the emission is effectively transposed to the power stations which are responsible for producing the electricity. Emissions from electric trains are therefore normally allocated to the power generation sector. Compared with road vehicles, trains generally have low emissions

per passenger-kilometer although this depends, of course, on the passenger load factors.

Shipping is responsible for a significant fraction of NO<sub>x</sub> and primary PM emissions in some countries. Many ships have old engines or old technology, and compared with road transport few exhaust aftertreatment devices are in use. The quality of the fuel is also poor relative to that used for road vehicles. In particular, the

residual fuel oil used in ship engines contains very high levels of sulfur (up to 4.5% by mass), which leads to high PM emissions [34]. To put this in perspective, the PM emission factor for ship engines is typically around 1.25 g/kWh [35], compared with the legislative limit value of 0.02 g/kWh for a modern Euro V truck. There is little information on the size distribution of PM emissions from ships, although some recent data indicate that a large fraction of the emitted particles can be classified as PM<sub>2.5</sub> [34]. NO<sub>x</sub> and sulfur dioxide (SO<sub>2</sub>) emissions from shipping also contribute significantly to secondary particle formation.

### Evaporation

Volatile organic compounds (VOCs) are emitted from the fuel systems of petrol vehicles as a result of evaporation. The compounds that are emitted are mainly of light hydrocarbons (C<sub>4</sub>–C<sub>6</sub>) [36]. Evaporative emissions from diesel-fuelled vehicles are considered to be negligible due to the low volatility of diesel fuel.

There are several different mechanisms of evaporation. “Diurnal losses” are due to the thermal expansion and emission of vapor, mainly in the fuel tank, as a result of changes in ambient temperature during the day. “Hot-soak losses” occur when a warm engine is turned off and heat is dissipated into the fuel system. Whilst a vehicle is being driven, the engine provides a continuous input of heat to the fuel system, resulting in “running losses.” In addition, “refueling losses” occur while the tank is being filled and the saturated vapors are displaced and vented into the atmosphere, but these are usually attributed to the fuel handling chain.

Evaporative emissions are dependent upon four major factors: the vehicle design, the ambient temperature, the volatility of the petrol, and the driving conditions. Emissions have decreased significantly in recent years as a result of new cars being equipped with sealed fuel injection systems and activated carbon canisters in fuel tank vents [5].

### Abrasion

As noted earlier, particles are also generated as a result of various unregulated abrasion processes, including tire wear, brake wear, and road surface wear. These

sources are becoming increasingly important (Fig. 4), but the understanding of emissions of abrasion particles is relatively poor and their health effects are not well documented.

There is little information on the composition of tire wear particles. Components that have been suggested as environmental tracers of tire wear include styrene butadiene rubber [38], benzothiazole [39], and organic zinc [40]. Studies conducted in the 1970s revealed that wear particles had a bimodal distribution, with mass in the both the fine (<2.5 μm) and coarse (2.5–10 μm) modes [38, 41–43]. Several laboratory and field investigations have demonstrated that most particles are present in the coarse fraction [38, 44, 45], or are larger still [46], whereas others have indicated substantial mass in both the PM<sub>2.5</sub> and PM<sub>1</sub> fractions [40, 47, 48]. In one early study, it was observed that mean particle diameter decreases with increasing speed, and this may be one factor contributing to the differences in the reported findings [42].

Airborne brake particles are composed of metals (iron, copper, lead, etc.), organic material, and silicon compounds, which are used as binders in brake pads, but composition varies greatly [37]. Test track and wind tunnel measurements have revealed that typically 50% of the brake wear debris escapes the vehicle and enters the atmosphere, although the actual proportion depends on the severity of the braking and the design of the vehicle [49]. The remaining debris either falls on the road or is retained on the vehicle. It appears that most airborne brake wear particles can be classified as PM<sub>10</sub>, although a substantial proportion has a diameter of less than 2.5 μm [49–52]. The highest proportions of PM<sub>2.5</sub> have been found under urban driving conditions [52].

Very little information is available on the size distribution of road surface wear particles. The data that are available suggest that airborne bitumen particles range in size between 0.35 μm and 2.8 μm [40].

Particle measurements conducted near a busy electrified railway line in Zurich showed that the main elements abraded by railway traffic were iron and manganese originating from the wheel–track interface, as well as copper from the overhead catenary system [53]. More than half of the iron and manganese particles resulting from wheel and track abrasion were in the

coarse range. The abrasion particles provided a small but measureable contribution to local PM<sub>10</sub> concentrations [54]. However, the contribution of rail transport on lines with a larger diesel share of the traction could be much higher.

### Resuspension

Another source of PM – the resuspension of material previously deposited on the road surface – occurs as a result of tire shear, vehicle-generated turbulence, and the action of the wind. Studies in the United States have indicated that resuspension is responsible for between 30% and 70% of total PM<sub>10</sub> in urban areas [55–57]. Large contributions of resuspension have also been observed in Germany [44, 58, 59] and the UK [60], albeit with a large variation between locations.

Resuspension is a major concern in Scandinavian countries, especially in relation to the use of studded tires and “road sanding.” Indeed, it can provide a much larger contribution to PM<sub>10</sub> than exhaust emissions, and also form a significant fraction of PM<sub>2.5</sub> [61, 62]. An investigation in Sweden revealed that PM<sub>10</sub> resuspension is highly dependent on vehicle speed [63].

### Modelling Transport Emissions

Estimates of emissions from transport are required for various purposes. National inventories covering all sectors are part of the reporting obligations of CLRTAP, the United Nations Framework Convention on Climate Change and, in the European Union, Directive 2001/81/EC on National Emission Ceilings. At a more local level, emission estimates are required for assessing the impacts of policies and measures, and for input into air quality modelling studies. Separate modelling approaches have been developed for the different transport modes. Most emission models are empirical in nature, being based on data from laboratory or real-world tests. However, it is worth noting that such tests are not always designed to deliver emission factors for modelling, and simulate only a small fraction of all possible operational conditions [5].

### Road Transport Models

**Combustion** In the simplest terms, the total emission of a given pollutant during a given time period is

calculated as the sum product of specific emission factors and quantities of activity:

$$E_p = \sum_{v=1}^{v=n} e_{p,v} \times T_v \quad (1)$$

where

$E_p$  = total emission (e.g., in grams) of pollutant  $p$  during the period

$v$  = vehicle category (with  $n$  different categories)

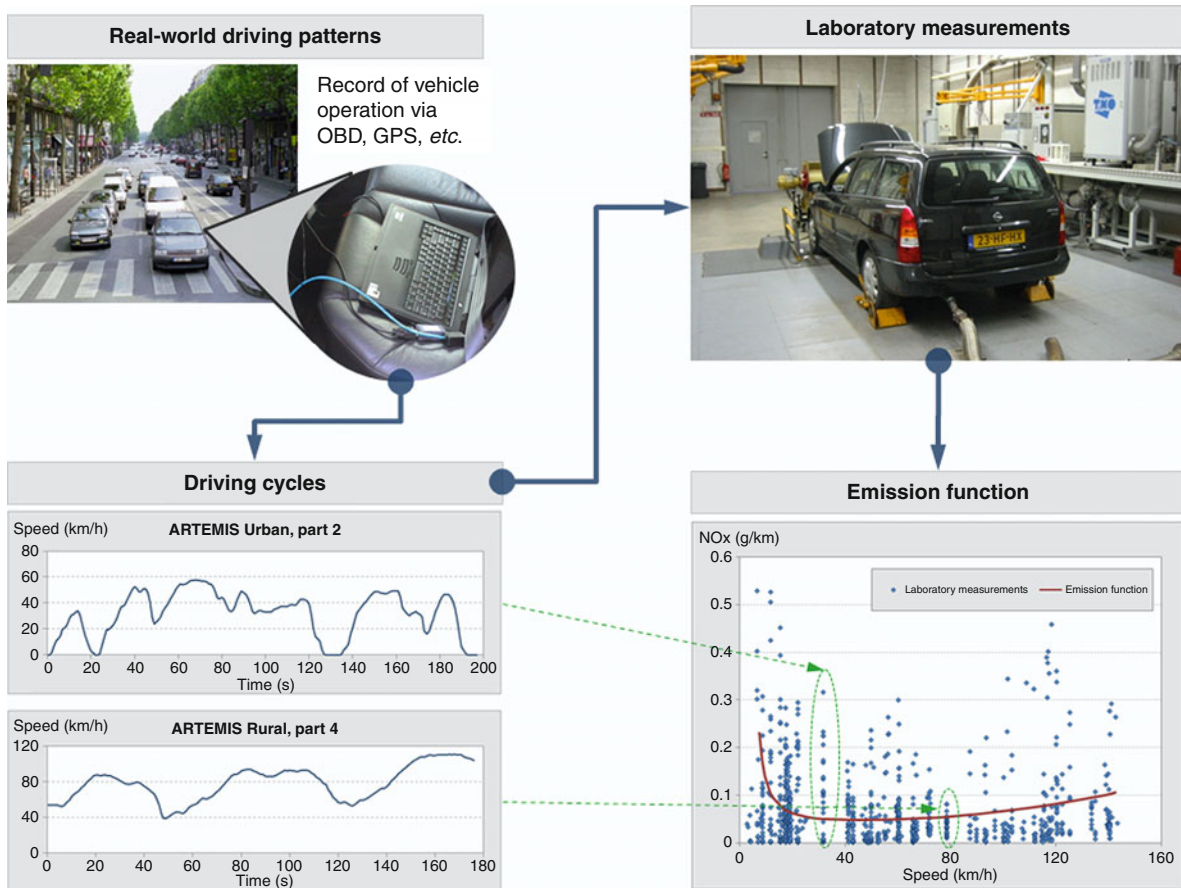
$e_{p,v}$  = emission factor (e.g., in grams per kilometer or liter of fuel burnt) of pollutant  $p$  from vehicle category  $v$

$T_v$  = activity (e.g., in kilometers or liters of fuel) of vehicle category  $v$

The definition of emission factors that represent different aspects of vehicles and their operation forms a substantial area of research. Emission models are generally classified in terms of how they take vehicle operation into account. Here, the term “vehicle operation” is used in a general sense, and refers to a wide range of parameters which describe the way in which a driver controls a vehicle (e.g., average speed, maximum speed, acceleration pattern, gear-change pattern), as well as the way in which the vehicle responds (e.g., engine speed, engine load). The main types of approach are aggregated emission factors, average-speed models, traffic situation models, multiple-regression models, modal models, and instantaneous models.

Aggregated emission factors operate on the simplest level, with a single emission factor representing a general type of driving; the traditional distinction is between urban roads, rural roads, and motorways. Vehicle operation is therefore only taken into account at a very rudimentary level. However, they are generally the only way of estimating emissions of unregulated pollutants, for which there is often insufficient information to define a more detailed relationship with vehicle operation.

Average-speed models are based upon the principle that the emission factor varies according to the average speed during a trip. The general process of model development is shown in Fig. 8. Firstly, vehicle operation in the real world is measured using on-board instruments, with the vehicles being driven under



**Air Quality, Surface Transportation Impacts on. Figure 8**

Process of development for average-speed models (Photographs reproduced by kind permission of Alistair Smith, Tim Barlow of TRL in the UK, and TNO Automotive in the Netherlands)

a wide range of road and traffic conditions. The numerous resulting driving patterns are then condensed, using statistical methods, into a much smaller number of representative driving cycles which can be used to test vehicles in the laboratory. The results from the laboratory tests on vehicles in the same category (e.g., small petrol cars complying with the Euro 4 legislation) are then plotted as a function of average speed, and a regression curve is fitted to the data to give an emission function. Given that exhaust emissions are inherently variable, there is usually a considerable amount of scatter in the emission data.

Average-speed models are widely used in emission inventories and local air pollution models. In Europe, the most widely used example is COPERT [33]. A number of factors have contributed to their

popularity. The modelling approach is one of the oldest, it is comparatively easy to use, several models are available free of charge, and there is a reasonably close correspondence between the required inputs and the data available to users. However, trips having very different vehicle operation (and different emission levels) can have the same average speed. Clearly, all the types of operation associated with a given average speed are not accounted for by the use of a single emission factor. This is a particular problem at low-medium average speeds, for which the range of possible operational conditions associated with a given average speed is great. In addition, for modern vehicles a large proportion of the total emission during a trip can be emitted as very short, sharp peaks, often occurring during gear changes and periods of high acceleration,



and this detail is lost when only average speed is considered. The concept of “driving dynamics” has therefore been developed to enable model developers to describe vehicle operation using additional parameters [64]. In qualitative terms, dynamics can, perhaps, be thought of as the “aggressiveness” of driving. Quantitatively, the term refers to the variation in various properties or statistical descriptors of a driving pattern. One of the more useful parameters appears to be relative positive acceleration (RPA) [65]. However, most model users have no straightforward means of quantifying such parameters.

In “traffic situation” models, emission factors are referenced to specific traffic situations that are known by the model user. A typical example is the Handbook of Emission Factors [66]. The traffic situations are characterized by the features of the sections of road concerned (e.g., “motorway with 120 km/h limit,” “main road outside built-up area”). The driving dynamics are not quantified by the user, but are defined through a textual description (e.g., “free flow,” “stop and go”). However, requiring the user to define the traffic situation using a textual description of dynamics may lead to inconsistencies in interpretation.

Multiple-regression models are another comparatively recent development, and the most prominent example is the Dutch VERSIT+ model [67]. The model contains test data for a large number of cars over a range of different driving cycles. Within the model, each driving cycle is characterized by various descriptive parameters (e.g., average speed, RPA, number of stops per km, etc.). A regression model is fitted to the average emission factors and cycle parameters for the various driving cycles, giving the descriptive variables which are the best predictors of emissions (the group of descriptors being different for each type of vehicle and each pollutant). The model requires a driving pattern (speed-time profile) as the input, from which it calculates the same range of descriptive variables and estimates emissions based on the regression results.

In modal models emission factors are allocated to the specific modes of vehicle operation encountered during a trip. In the simplest type, vehicle operation is defined in terms of a relatively small number of modes – typically idle, acceleration, deceleration, and cruise. For each of the modes the emission rate is fixed,

and the total emission during a trip, or on a section of road, is calculated by weighting each modal emission rate by the time spent in the mode [68, 69]. This approach has usually been used to determine the impacts of traffic control measures and signal improvements [70].

Some more detailed modal models aim to give a precise description of vehicle emission behavior during a series of short time steps (often 1 s). These are commonly referred to as “instantaneous” models, although several alternative terms can be found in the literature, including “microscale,” “continuous,” and “online” [71]. In the simplest type of instantaneous model, emission rates are defined for different combinations of speed and acceleration, usually in the form of a matrix [72, 73]. Theoretically, the finer the resolution of the emission factor matrix the higher the model accuracy, but as the resolution improves the complexity of the calculations increases. Some models use the product of speed and acceleration instead of the acceleration alone [74]. A more flexible approach relies upon emission maps that relate to engine power [75, 76]. This allows all relevant factors to be taken into account, including rolling resistance, aerodynamic resistance, acceleration, gradient, transmission losses, and the use of auxiliaries.

Instantaneous models should have a number of advantages. Firstly, emissions can be calculated for any operational profile, and thus new emission factors can be generated without the need for further vehicle testing. Secondly, the models inherently take into account dynamics. Thirdly, emissions can be resolved spatially, and this ought to lead to improvements in the prediction of air pollution. However, detailed and precise data on vehicle operation and location are required, or any potential benefits may be lost. Such data are difficult to obtain for many model users. Moreover, it is extremely difficult to measure emissions on a continuous basis with a high degree of precision, and then it is not straightforward to allocate those emission values to the correct operating conditions. During measurement in the laboratory, an emission signal is dynamically delayed and smoothed, and this makes it difficult to align the emissions signal with the vehicle operating conditions [77]. Such distortions have not been fully taken into account in instantaneous models until relatively recently.

Once the basic modeling approach has been identified a number of other effects may need to be considered, depending on the type of model being used. These effects include potential future improvements to fuels and technologies, the deterioration of vehicle exhaust emissions with age, the presence and use of auxiliary devices (e.g., air conditioning, windscreen wipers, radios), road gradient, and vehicle load/occupancy. Such effects may be taken into account via separate models [78] or through adjustments to the basic emission factors, with assumptions concerning levels of uptake and effectiveness in future years.

Separate models also exist for the estimation of cold-start emissions. In some cases, the hot emission factor is multiplied by a ratio of cold-start to hot emissions for the fraction of kilometers driven with a cold engine [33]. In others, a reference cold-start excess emission value is adjusted to take account of, for example, the ambient temperature, the average speed, the parking duration, and the distance travelled (some trips are shorter than the distance needed to fully warm the engine) [79, 80]. Specific models have also been developed to determine the effects of different parking scenarios [81]. Cold-start emissions are usually included in regional and national emission inventories, but are often excluded from local air pollution studies. One of the main reasons for this is the difficulty associated with obtaining data on the temporal and spatial allocation of cold-start operation.

**Evaporation** The earliest studies of evaporative emissions were carried out in the United States in the 1960s [82], but the first European studies were only conducted 20 years later [36]. The models that were developed from the resulting data were not updated for many years. However, modelling improvements have recently been possible following a number of new investigations [83–85].

**Abrasion and Resuspension** One non-exhaust PM model in common use is that presented in the European Environment Agency's Air Pollutant Emissions Inventory Guidebook [33]. The method uses reference emission factors (in g/km) and incorporates corrections for both speed and, in the case of HDVs, vehicle load and number of axles. However, the

use of average brake wear emission factors (in g/km) means that differences in the extent of braking for different traffic situations cannot be taken into account. Such emission factors are likely to overestimate brake emissions in areas of low brake use, and underestimate emissions in areas of high brake use.

The calculation method presented in the US Environmental Protection Agency's AP-42 document [86] is one of the more widely used for predicting resuspension on paved roads, although some shortcomings have been identified [87]. There is no widely accepted European method for calculating emissions due to resuspension.

### Rail Transport Models

Emissions from electric trains occur at the power stations which are responsible for generating the electricity. The actual emissions will therefore be a simple function of power consumption and emission factors for power generation. However, power losses between the power station and the train, and within the train itself, also need to be taken into account where possible.

Emissions from diesel locomotives are typically estimated using fuel consumption data and fuel-specific emission factors. The types of train which are typically included in models include high-speed passenger trains, intercity passenger trains, regional passenger trains, urban passenger trains, and goods trains. Early models were based on empirical correlations between emissions, train type, operating speed, and other simple variables such as the distance between stops, and average emission factors for train types [88, 89]. More recent models allow the user to specify an operational matrix for a journey, from which energy consumption and emissions can be calculated [90, 91]. This approach is a broadly analogous to that used in modal models for road transport. As with road vehicles, passenger trains consume energy for loads other than propulsion. These include the heating and cooling loads, and electrical power to operate lights, instruments, and other electrical accessories. However, the energy consumption for auxiliary power is fairly low compared with that required for propulsion. The effects of regenerative braking on energy consumption may also be important.

## Shipping Models

For shipping, both simple and detailed methods have been developed for calculating emissions [33, 92, 93]. The simple methods are essentially similar to that for road transport (see Eq. 1), and combine fuel use (in tonnes) with fuel-specific emission factors. Detailed calculation approaches firstly involve obtaining ship movement data, including sailing routes and distances between ports. Different ship operating modes may also be considered, such as cruising, in-port operation (hotelling, loading and unloading), and maneuvering. Each vessel also needs to be characterized. National and international shipping covers a wide range of vessels, from recreational craft to large oceangoing cargo ships that are driven primarily by large, low-, and medium-speed diesel engines. In addition, marine engines can be further classified according to their principal fuel: bunker fuel oil, marine diesel oil, or marine gas oil. Again, auxiliary engines are used to provide power and services when berthed [33]. In Europe, a widely used set of emission factors was determined in a study for the European Commission [92].

The fuel consumption and emissions of vessels operating on inland waterways were initially modelled by analogy to marine shipping. However, the operational characteristics of inland ships are rather different from those of their seagoing counterparts. Firstly, they sail at lower speeds and in restricted waterways. The lower speeds reduce energy consumption and emissions, whereas shallower depths increase sailing resistance and hence fuel consumption and emissions. In order to make the most of the limited space on inland waterways, vessels tend to be squarer in shape than seagoing ships. This also causes the sailing resistance to differ from that of seagoing ships [94]. Recent models are therefore more specific to inland vessels and their operating conditions. For example, in the ARTEMIS project, a model was developed and validated for nine types of vessel based on size and type of waterway [94]. The model calculates the energy consumption and emissions under different service conditions (speed, waterway dimensions, load, etc).

## Activity Data

For all transport modes the accurate characterization of traffic activity (such as numbers vehicles, trip distances,

and modes of operation) and the fleet composition is vital to the estimation of emissions. Although models and emission factors are continually improving, poor activity data remains one of the main sources of uncertainty in the calculation of emissions [95]. For many years, the direct measurement of road traffic has been the primary source of activity data. The collection of accurate information on important parameters such as traffic speed, road gradient, or vehicle load can often be difficult and time-consuming, rendering it impractical for large geographic areas. However, the quality and availability of activity data has increased substantially in recent years as a result of improved data collection technologies (e.g., tag-and-beacon systems, cameras, in-vehicle data collection, and satellite technology) and online dissemination. Furthermore, there have been significant advances in microsimulation traffic models, the outputs from which are well suited to integration with instantaneous emission data. Some workers have linked microscale traffic and emission models for practical applications [96].

In the case of passenger rail transport the activity data are available from timetables. Rail freight movements are more complicated, and data are generally more difficult to obtain. There are essentially three types of freight movement: long-term scheduled, short-term scheduled, and “on demand.” Clearly, it would be difficult to incorporate on-demand movements in any emission calculations. Encouragingly, an increasing number of operators are adopting “black box” technologies, which can provide useful journey and fleet information. For example, in Germany, the rail company Deutsche Bahn has installed equipment to monitor the real-time location and energy consumption of its locomotives [97]. However, at the moment, it appears that access to the data from train event recorders is rather limited.

The primary source of information on ship movements is the database provided by Lloyd’s Maritime Information Service. The database includes ship size, destination, approximate time of arrival and departure, engine type and number, but only for ships greater than 500 gross tons. Ferries, fishing vessels, and operations at smaller ports are not included. Port calling statistics are also available from national sources. In the future, the availability and accuracy of activity data and emissions data should improve further due to onboard GPS

technology, continuous emission monitoring systems (CEMS) and the development of a universal Automatic Identification System (AIS) [98]. For inland shipping there is currently a need for data that give a better description of the fleet, the traffic, and the waterways.

### Pollutant Dispersion and Transformation

Once pollutants have been released into the atmosphere, they are subject to various processes of dispersion and physical or chemical transformation. Emissions from transport are also superimposed on emissions from other sources. Figure 9 shows a simplified representation of pollutant concentrations in and around an urban area with tall commercial buildings in the urban center and low residential buildings in the surrounding districts. The regional background pollution originates from a range of sources, extends over a wide area, and is relatively invariant outside the urban area. Within the built-up area there is an additional “urban background” component. This results from combined emission sources located within the urban area, including transport, power plant, industry, and domestic heating. Alongside heavily trafficked roads there is also likely to be a significant local contribution to the concentration.

### Dispersion

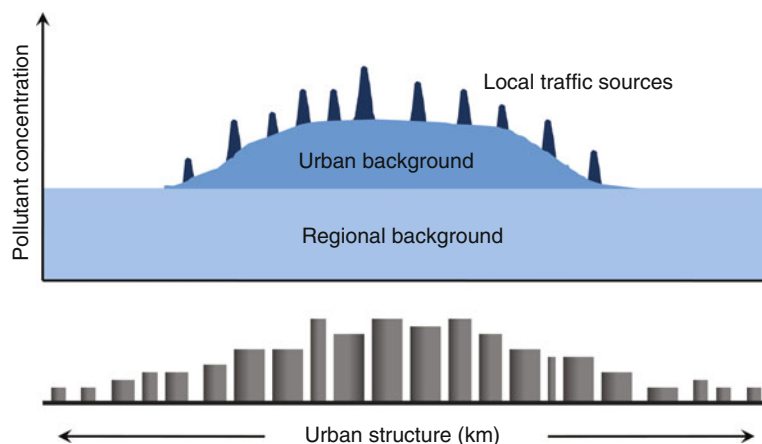
The dispersion of pollutants is influenced to a large extent by the local weather conditions (see entry on

► [Urban Air Quality: Meteorological Processes](#)). In particular, the temperature inversions and low wind speeds associated with stable high-pressure systems can restrict dispersion and lead to high pollutant concentrations. Dispersion is also influenced by the local topography, and is reduced by the presence of local obstacles such as buildings. The effect is most pronounced in the case of so-called street canyons. Moreover, vehicle-induced turbulence needs to be taken into account.

### Transformation

Some of the transformation processes that influence the concentrations of PM, NO<sub>2</sub> and ozone are summarized below. Meteorological factors, including solar radiation, precipitation, and humidity, also play important roles in these processes.

**Particulate Matter** The emission of primary particles from combustion sources was described earlier. Nucleation mode particles have a relatively short existence in the atmosphere since they readily transform into larger particles and deposit quite efficiently to surfaces because of their highly diffusive nature. Accumulation mode particles are too large to be subject to rapid diffusion and too small to settle from the air rapidly under gravity. Their further growth is inhibited because they do not coagulate quickly and there are diffusion barriers to their growth by condensation. Particles in



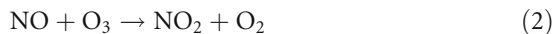
**Air Quality, Surface Transportation Impacts on. Figure 9**

Simplified representation of urban structure and pollution levels (Adapted from [18])

the accumulation mode can therefore have a long atmospheric lifetime (typically 7–30 days), although incorporation into rain can shorten it significantly. Within the coarse particle size range, gravitational settling velocities become appreciable and therefore atmospheric lifetimes are much shorter than for accumulation mode particles. Particles larger than  $\sim 100\ \mu\text{m}$  in diameter rapidly settle out of the air and are of minor health significance because, although they can be inhaled, they do not generally penetrate beyond the nose and mouth [99].

A substantial fraction of the fine PM mass, especially at background locations, is secondary in nature [99–101]. Secondary particles are formed by atmospheric reactions involving both inorganic and organic gaseous precursors, several of which are emitted by transport. For example, particles can be produced by the oxidation of precursor gases such as  $\text{SO}_2$  and  $\text{NO}_x$  to acids followed by neutralization with ammonia ( $\text{NH}_3$ ), or by the partial oxidation of organic compounds. The formation of secondary particles is relatively slow. Their persistence in the atmosphere is therefore prolonged, and they are distributed more evenly throughout the air than primary particles. They can also travel large distances, resulting in the transport across national boundaries.

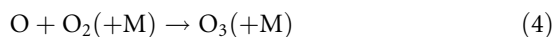
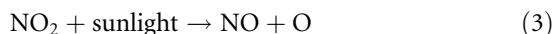
**Nitrogen Dioxide** The ambient concentration of  $\text{NO}_2$  is dictated by various complex processes and factors. Under the majority of atmospheric conditions the following reaction is the dominant pathway by which NO is converted to  $\text{NO}_2$ :



At polluted locations comparatively close to sources of  $\text{NO}_x$  (such as road vehicles), NO is in large excess and the reaction proceeds quickly (seconds), with the availability of  $\text{O}_3$  limiting the quantity of  $\text{NO}_2$  that can be produced. Primary  $\text{NO}_2$  emissions are therefore important at such locations. At unpolluted locations, when  $\text{O}_3$  is present in excess, the timescale for conversion of NO to  $\text{NO}_2$  is longer (minutes). When neither NO nor  $\text{O}_3$  is in large excess, the reaction progressively depletes both reagents, and can become very slow [20].

Under normal ambient daytime conditions, the reverse process also occurs – the destruction of  $\text{NO}_2$

by photolysis to form NO and ozone, as shown in reactions (3) and (4):



where M is a third body, most commonly nitrogen.

The mean daylight lifetime of  $\text{NO}_2$  with respect to photolysis is around 3 min. In winter, it is typically a factor of 2 or 3 longer. Under conditions when photolysis is sufficiently rapid, NO,  $\text{NO}_2$ , and  $\text{O}_3$  are potentially in chemical equilibrium, a condition usually referred to as the “photostationary state.”

Other chemical processes that can convert NO to  $\text{NO}_2$  involve the formation of free radicals, which can catalyze the oxidation of emitted hydrocarbons [20], but these reactions tend to be comparatively slow [102].

A third-order reaction, involving the oxidation of NO by oxygen, is of minor importance under normal atmospheric conditions:



However, this reaction is rapid at low temperatures and high NO concentrations, such that it is much more rapid at the elevated levels typical of those close to points of emission. Under wintertime pollution episode conditions, when a shallow inversion layer can lead to a combination of high  $\text{NO}_x$  levels and stagnant air for periods of a day or more, reaction (5) can potentially make a substantial contribution to  $\text{NO}_2$  concentrations, as during an episode in London in December 1991 [20].

Given that model predictions show a decrease in  $\text{NO}_x$  emissions, the absence of a downward trend in  $\text{NO}_2$  concentrations (see earlier section on “Historical trends”) is not completely understood. Three contributing factors have often been cited:

1. The likelihood that the average proportion of  $\text{NO}_2$  in  $\text{NO}_x$  (termed  $f\text{-NO}_2$ ) in diesel exhaust is increasing. Analyses of  $\text{NO}_2$  concentrations in urban areas have indicated that a significant proportion of  $\text{NO}_2$  must be emitted directly from the exhaust, and that the  $f\text{-NO}_2$  value has increased in recent years [104–108]. This has two causes:
  - The increasing market share for diesel vehicles in many European countries. For example, the



share of first registrations of diesel passenger cars in Finland increased from 17% in 2005 to 52% in 2008 [109]. Petrol cars emit less  $\text{NO}_x$  than diesel vehicles, and with a smaller proportion of  $\text{NO}_2$ .

- The increase in the use of specific aftertreatment technologies in modern diesel vehicles which involve in situ generation of  $\text{NO}_2$ , such as catalytically regenerative DPFs [106].

The changes in  $f\text{-NO}_2$  for traffic emissions have not had a large impact on  $\text{NO}_2$  concentrations away from the roadside, where the dominant contribution is from secondary  $\text{NO}_2$ .

2. The possibility that real-world  $\text{NO}_x$  emissions from road vehicles are not decreasing as rapidly as models are predicting [110].
3. Increasing background ozone concentrations [111]. As the  $\text{O}_3$  concentration increases, the amount of  $\text{NO}$  converted to  $\text{NO}_2$  increases.

**Ozone** Ground-level  $\text{O}_3$  is implicated in summer smog, such as the one that affected large areas of Europe in the summer of 2003, and high concentrations lead to increases in the frequency of respiratory symptoms and in deaths [112]. It is not produced directly from emission sources but is created by photochemical reactions such as those described above. It can be transported over long distances, and is therefore regarded as a regional air pollution problem. High concentrations are typically observed downwind of large cities in the summer when photochemical formation is enhanced. Because road transport is a major source of ozone precursors (e.g.,  $\text{NO}_x$  and hydrocarbons), it is an important contributor to ground-level concentrations. Globally, road transport is responsible for 5–15% of tropospheric  $\text{O}_3$  [113, 114].

### Modelling Air Pollution from Transport

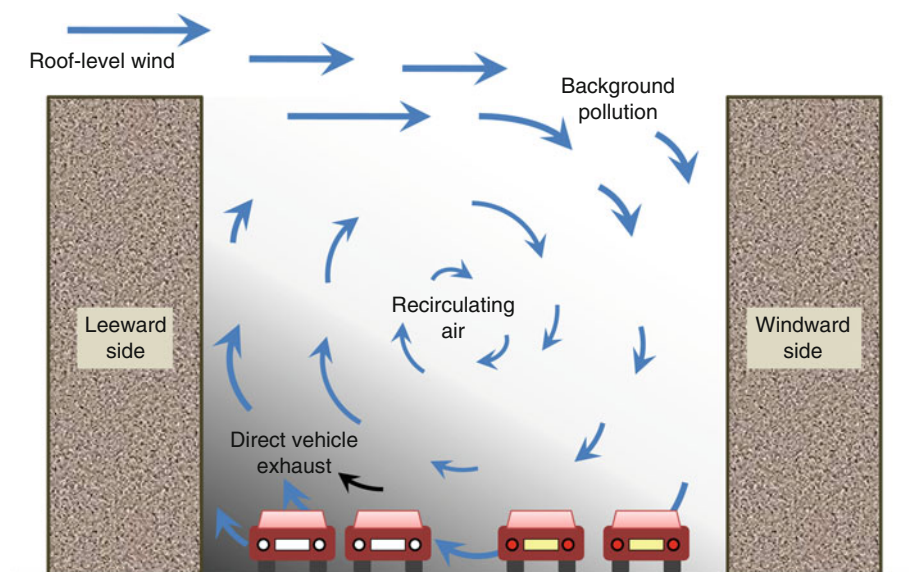
Predictions from air quality models are used in a diverse range of applications. A common one is air quality management (AQM), which includes the identification of high-pollution areas (“hot spots”), the development of plans and programs for improving air quality, the checking of progress toward targets, and the provision of information to the public. Applications that are closely linked to AQM are transport planning

and land-use planning. Air quality predictions are also an important element of environmental impact assessment (EIA).

Modelling methodologies can be categorized according to two general types: empirical/statistical models and deterministic models. Empirical models are based largely on correlations between pollutant concentrations and other variables, such as meteorological parameters. Although empirical models can be used for short-term predictions, they are most reliable for longer term averages. The assumptions that correlations represent cause and effect, and that regression parameters remain constant into the future, are their main weaknesses [20]. Deterministic models predict concentrations downwind of an emission source using mathematical formulae which describe the physical and chemical processes operating within the atmosphere [20]. They include Gaussian plume models, Eulerian grid models, Lagrangian trajectory models, and computational fluid dynamics (CFD) models. For transport sources, emissions are typically entered as a line source (such as a road or railway line), or as an area/volume source (such as a car park).

Models vary enormously in sophistication. The simplest and most common deterministic models are based on Gaussian dispersion theory. This assumes that a pollutant emission develops into a plume under the influence of the wind, and that the concentrations in the plume have Gaussian distributions in the horizontal and vertical directions. Two other types of model – Eulerian and Lagrangian – take as their basis the principle of conservation of mass of a pollutant as it spreads in the air. In Eulerian models, the calculations are undertaken simultaneously for all the grid points in the model domain, whereas in the Lagrangian approach the calculations relate to a parcel of air which follows a defined trajectory. Several thousand trajectories are usually required to generate statistically significant results. CFD models use numerical methods to predict air flows. They focus on the detailed modelling of complex structures on the local scale (e.g., street canyons), and are not appropriate for calculating area-wide concentrations [20].

The most commonly used street canyon model is the Operational Street Pollution Model (OSPM) [115]. The OSPM approach involves an idealized



**Air Quality, Surface Transportation Impacts on. Figure 10**  
Schematic illustration of street canyon modelling in OSPM (Adapted from [115])

representation of a street as a long road enclosed along its length on both sides by buildings of equal height (Fig. 10). The wind passing across the street induces a recirculation vortex in the street, and this leads to elevated pollutant concentrations on the leeward side of the canyon.

Some models include a preprocessor module for the input of meteorological data. Allowances can also be made for concentration changes produced by chemical reactions in the air. Again, the chemical reaction schemes vary greatly in sophistication.

Dispersion model outputs are often expressed as contour or surface maps in order to show the spatial variation in pollutant concentrations over a wide area. Such plots can enable spatial relationships between air pollution and other variables to be established, and sensitive locations to be identified. Appropriate pollution-reduction measures can then be developed. The example provided in Fig. 11 shows  $\text{NO}_2$  concentrations in London. The area with high  $\text{NO}_2$  concentrations to the west of London is Heathrow Airport.

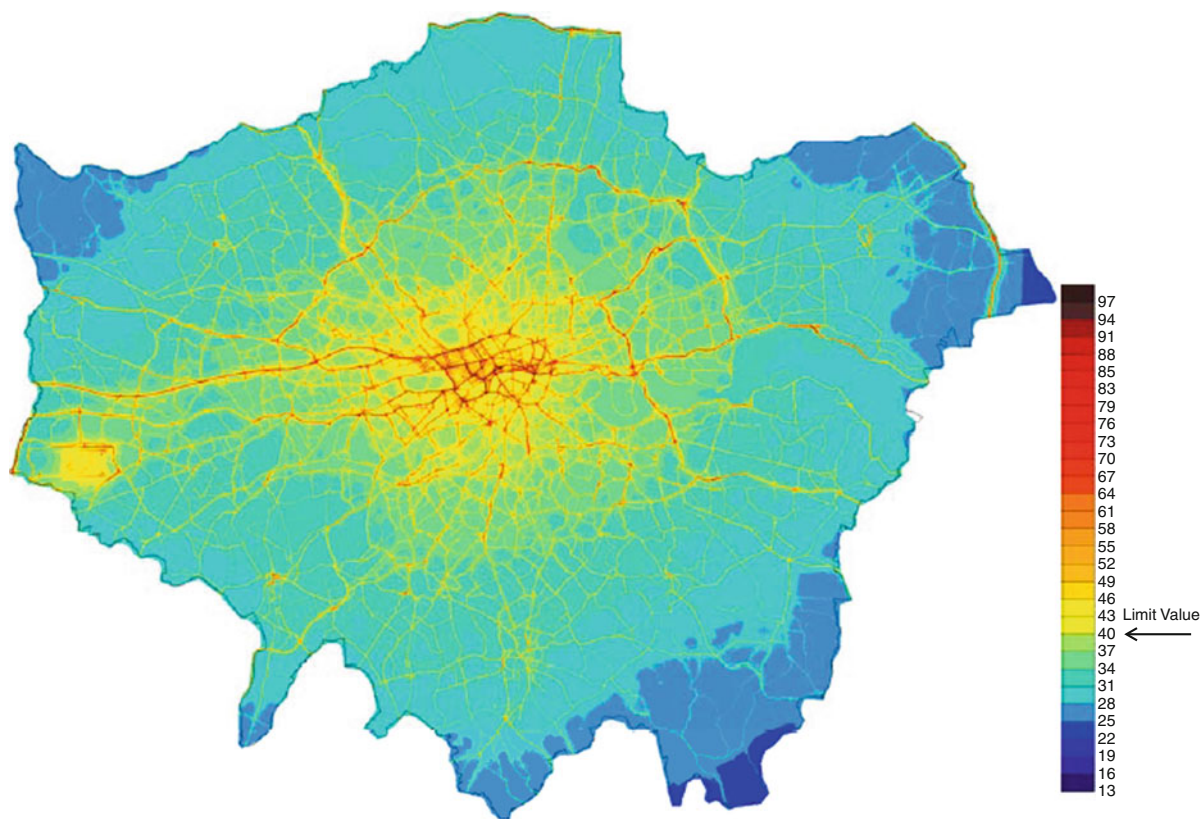
### Reducing Air Pollution from Transport

If transport systems are to become environmentally sustainable their impacts on air pollution will have to

be greatly reduced, and it is therefore important to consider ways in which this might be achieved. The emphasis here is on road transport, but some measures for rail transport and shipping are also mentioned. It has been noted that if no aggressive emission-reduction strategies are introduced, global  $\text{CO}_2$  and  $\text{SO}_2$  emissions from ships could double from present-day values by 2050, and  $\text{NO}_x$  emissions could exceed the current value for road transport [117]. Shipping has therefore been recognized as a growing problem by policymakers and scientists.

This entry considers two philosophies: “prevention” and “mitigation.” Of course, other distinctions are possible (e.g., “technical,” “systematic,” “behavioral,” etc.), but in any case these tend to be artificial; interactions are often so strong that few strategies are based on a single measure of policy. In the context of the entry, prevention measures are designed to reduce or eliminate emissions at source, whereas mitigation measures are designed to remove pollutants which have already been emitted, or to convert them to more benign compounds.

An important tool is air quality legislation (see entry on ► [Air Quality Guidelines and Standards](#)), but it does not fall conveniently into the prevention or mitigation categories. Although air quality



**Air Quality, Surface Transportation Impacts on. Figure 11**

Modelled annual mean NO<sub>2</sub> concentrations (µg/m<sup>3</sup>) in London in 2008 [116] (Photograph reproduced by kind permission of Greater London Authority)

legislation acts as a driver for the development of pollution-reduction policies and technologies, and can thus be considered as “prevention,” compliance with the legislation can only be determined from historical data, usually for the calendar year, such as the annual mean or the number of days on which pollutant concentrations were above a certain value. Neither does the legislation specifically target transport.

Compliance with air quality standards for PM<sub>10</sub> requires control of both fine and coarse particles. As the two modes tend to have different sources and formation mechanisms, different types of control are required. The compounds that are precursors to secondary particles also need to be controlled. Primary fine particles from combustion sources are subject to regulation. The control of coarse particles is less straightforward [118], as such particles arise from diverse natural and anthropogenic disruption and

attrition processes which are difficult to characterize (e.g., non-exhaust emissions, resuspension, dust from industrial processes, quarrying).

As a result of reaction (2), O<sub>3</sub> levels near to roads tend to be low. Thus, the limitation of local traffic emissions would tend to lead to increased O<sub>3</sub> concentrations at roadside locations. The control of O<sub>3</sub> depends ultimately on the control of emissions of NO<sub>x</sub> and HC on a larger scale.

### Prevention Measures

Some examples of prevention (i.e., emission-reduction) measures are listed in Table 1. The list is not comprehensive, but provides an indication of the types of action which are currently being considered. Two general approaches are noted here: reducing emissions per vehicle and reducing travel/transport activity, and

Air Quality, Surface Transportation Impacts on. Table 1 Prevention measures for surface transport

| General approach                                | Type of measure           | Measure                       | Applicability |      |          |
|-------------------------------------------------|---------------------------|-------------------------------|---------------|------|----------|
|                                                 |                           |                               | Road          | Rail | Shipping |
| Reducing emissions per vehicle, vessel or train | Emission legislation      | Vehicle/engine type approval  | ✓             | ✓    | ✓        |
|                                                 |                           | In-service inspection         | ✓             | –    | –        |
|                                                 |                           | Roadside checks               | ✓             | –    | –        |
|                                                 | Fuel legislation          | Improved fuel quality         | ✓             | ✓    | ✓        |
|                                                 | Exhaust aftertreatment    | DPF                           | ✓             | ✓    | ✓        |
|                                                 |                           | SCR                           | ✓             | ✓    | ✓        |
|                                                 |                           | Scrubber                      | –             | –    | ✓        |
|                                                 | Alternative fuels         | CNG/LPG                       | ✓             | ✓    | ✓        |
|                                                 |                           | Biofuels (biodiesel, ethanol) | ✓             | ✓    | ✓        |
|                                                 | Vehicle/engine technology | Improved combustion           | ✓             | ✓    | ✓        |
|                                                 |                           | Improved aerodynamics         | ✓             | ✓    | ✓        |
|                                                 |                           | Hybrid                        | ✓             | ✓    | ✓        |
|                                                 |                           | Fuel cell                     | ✓             | ✓    | ✓        |
|                                                 |                           | Humid air motors              | –             | –    | ✓        |
|                                                 | Eco-driving               | Training                      | ✓             | ✓    | ✓        |
|                                                 |                           | Engine switch-off             | ✓             | ✓    | ✓        |
|                                                 | Cleaning                  | Road sweeping                 | ✓             | –    | –        |
|                                                 |                           | Road washing                  | ✓             | –    | –        |
|                                                 |                           | Vehicle cleaning              | ✓             | –    | –        |
|                                                 | Dust suppressants         | Chloride salts, CMA           | ✓             | –    | –        |
|                                                 | Traffic management        | Speed control                 | ✓             | –    | –        |
| Reducing travel/ transport activity             | Traffic management        | Access restrictions           | ✓             | –    | –        |
|                                                 |                           | Environmental zones           | ✓             | –    | –        |
|                                                 |                           | Parking restrictions          | ✓             | –    | –        |
|                                                 | Economic measures         | Congestion charges            | ✓             | –    | –        |
|                                                 |                           | Tolls                         | ✓             | –    | –        |
|                                                 |                           | Road user charging            | ✓             | –    | –        |
|                                                 |                           | Incentives and taxes          | ✓             | ✓    | ✓        |

several specific measures are available in each case. The applicability of each measure to road transport, rail transport, and shipping is also shown. Several prevention measures are described in more detail in the following paragraphs.

**Emission Legislation** In Europe, emission tests are required at type approval for all new light-duty vehicle models and for the engines used in heavy-duty vehicles. Light-duty vehicles are tested using a power-absorbing chassis dynamometer, as shown in Fig. 8, whereas

heavy-duty engines are operated on a test bench. The sampling of exhaust emissions is then performed as the vehicle or engine progresses through a predefined driving cycle.

Limit values are usually defined for CO, HC, NO<sub>x</sub>, and PM in exhaust, as well as evaporative emissions of VOCs. In the regulation of exhaust particles, practically all worldwide emission standards are expressed in terms of the mass of total particulate matter with no differentiation by particle size. Particle number emissions are included in the latest European legislation.

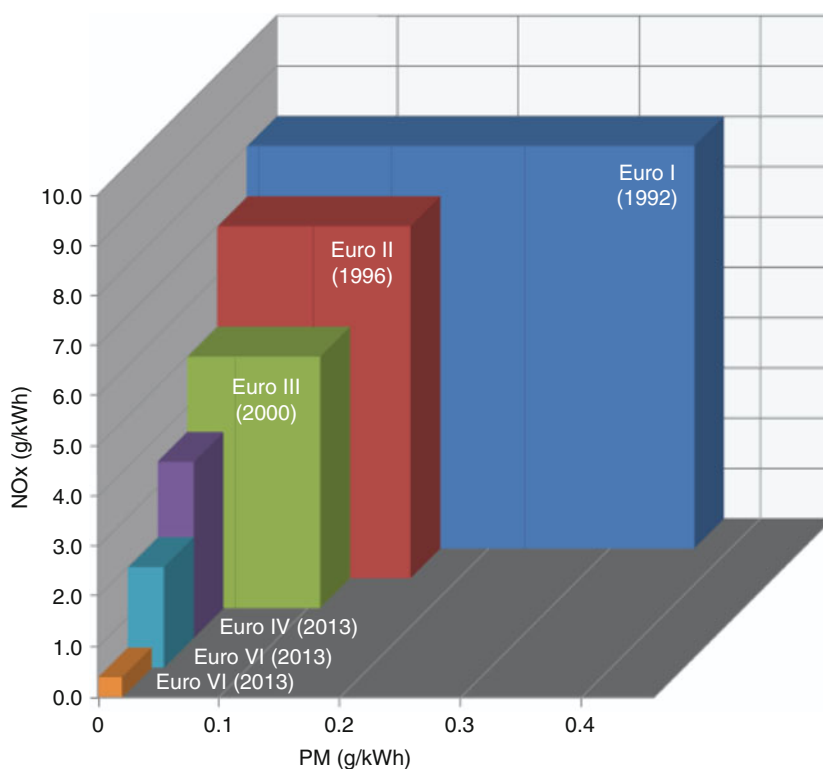
The most advanced regulations are to be found in the EU, the United States, and Japan. The specific differences between the legislation in different countries are too numerous to mention here, but the principles involved are broadly similar and there is a general move toward harmonization. Moreover, the technology being built into cars and trucks to comply with the European, US, and Japanese standards is very similar.

Over the course of several decades the limit values for the pollutants in vehicle exhaust have decreased

(Fig. 12), and changes have been made to the test method to make it more realistic. As a result, manufacturers have been required to develop increasingly effective emission-control technologies (some effects of this were shown in Fig. 4b). For example, the latest EU emission standards for cars and heavy-duty vehicles require large reductions in PM and NO<sub>x</sub> emissions, and these will be achieved through the use of DPFs and SCR.

Some examples of the progressive reduction in the exhaust emission limits in several countries (in terms of equivalence to the EU legislation) are illustrated in Fig. 13. Most developing countries lag behind the EU, although the gap is narrowing and in some places it has been possible to “leapfrog” standards. However, a prerequisite for the introduction of advanced engine and aftertreatment technologies is the availability of appropriate fuel, and in many developing countries this has presented something of a barrier.

Periodic in-service inspection and maintenance represents a further (and crucial) level of legislative



**Air Quality, Surface Transportation Impacts on. Figure 12**

Evolution of EU NO<sub>x</sub> and PM emission limits for engines used in heavy-duty vehicles



|                  | 1992   | 1993 | 1994   | 1995   | 1996   | 1997   | 1998   | 1999   | 2000   | 2001   | 2002   | 2003   | 2004   | 2005   | 2006 | 2007   | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|------------------|--------|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|------|--------|------|------|------|------|------|------|------|------|
| EU               | Euro 1 |      |        | Euro 2 |        |        | Euro 3 |        |        | Euro 4 |        |        | Euro 5 |        |      | Euro 6 |      |      |      |      |      |      |      |      |
| Thailand         |        |      | Euro 1 |        |        |        | Euro 2 |        | Euro 3 |        |        | Euro 4 |        |        |      |        |      |      |      |      |      |      |      |      |
| Singapore        |        |      | Euro 1 |        |        |        | Euro 2 |        |        |        |        |        |        |        |      |        |      |      |      |      |      |      |      |      |
| Hong Kong        |        |      | Euro 1 | Euro 2 |        |        | Euro 3 |        |        | Euro 4 |        |        | Euro 5 |        |      |        |      |      |      |      |      |      |      |      |
| Australia        |        |      |        | Euro 1 |        |        |        | Euro 2 |        | Euro 3 |        | Euro 4 |        |        |      |        |      |      |      |      |      |      |      |      |
| Malaysia         |        |      |        | Euro 1 |        |        |        |        |        | Euro 2 |        | Euro 4 |        |        |      |        |      |      |      |      |      |      |      |      |
| Vietnam          |        |      |        |        | Euro 1 |        |        |        | Euro 2 |        |        |        |        |        |      |        |      |      |      |      |      |      |      |      |
| Russia           |        |      |        |        | Euro 1 |        |        |        | Euro 2 |        | Euro 3 | Euro 4 |        | Euro 5 |      |        |      |      |      |      |      |      |      |      |
| Argentina        |        |      |        |        |        | Euro 1 |        | Euro 2 |        | Euro 3 | Euro 4 |        |        |        |      |        |      |      |      |      |      |      |      |      |
| China (national) |        |      |        |        |        | Euro 1 |        | Euro 2 |        | Euro 3 | Euro 4 | Euro 5 |        |        |      |        |      |      |      |      |      |      |      |      |
| India (national) |        |      |        |        |        | Euro 1 |        | Euro 2 |        |        | Euro 3 |        |        |        |      |        |      |      |      |      |      |      |      |      |
| Indonesia        |        |      |        |        |        | Euro 1 |        | Euro 2 |        |        |        | Euro 4 |        |        |      |        |      |      |      |      |      |      |      |      |
| Philippines      |        |      |        |        |        |        |        | Euro 1 |        | Euro 2 |        |        | Euro 4 |        |      |        |      |      |      |      |      |      |      |      |
| Pakistan         |        |      |        |        |        |        |        | Euro 1 |        |        |        |        |        |        |      |        |      |      |      |      |      |      |      |      |

**Air Quality, Surface Transportation Impacts on. Figure 13**

Introduction dates for emission standards in different countries (Adapted from [119])

control, as it is designed to ensure that vehicles maintain low emission levels during real-world operation. However, whilst type approval checks are relatively detailed and involve the use of specialist and expensive laboratory equipment, in-service checks are based on shorter, simplified operations of the vehicle (usually in a repair center), and make use of equipment that is less precise and less expensive than that used in the laboratory. Moreover, the tests for petrol vehicles typically only involve the measurement of CO and HC whilst the vehicle's engine is running at idle or at fast-idle, and those for diesel vehicles tend to cover only smoke emissions. Therefore, the tests do not address the most important pollutants and operational conditions. In some countries, attempts have been made to address these issues. For example, in the United States, a loaded test (IM240) has been developed to simulate conditions under which high NO<sub>x</sub> emissions may occur. Some authorities may also conduct random roadside checks, or use remote sensing devices to screen for high emitters. In addition, onboard diagnostics of emission-control systems are increasingly used to alert drivers to failures.

In Europe emissions from ships and railway locomotives are governed by the legislation for non-road

mobile machinery (Stages III and IV of Directive 2004/26/EC). The Stage III standards are further divided into two substages: Stage III A and Stage III B. The Stage III B standards introduce PM limit of 0.025 g/kWh, representing a 90% emission reduction relative to Stage II. To meet this limit value it is anticipated that engines will have to be equipped with DPFs. Stage IV also introduces a stringent NO<sub>x</sub> limit of 0.4 g/kWh, which is expected to require NO<sub>x</sub> after treatment. Given the long working lives of the engines in question, these regulations will not have any considerable effect on emission levels for another couple of decades [94]. Again, a process of worldwide harmonization of regulations has begun.

Emission standards for diesel engines used in railway locomotives have also been established by the International Union of Railways (UIC), an association of European railway companies [120]. The standards apply to all new engines used in new vehicles or for repowering of existing locomotives, and are binding to UIC members.

Emission standards for marine engines are defined by the International Maritime Organization (IMO). The standards – for NO<sub>x</sub> and SO<sub>x</sub> – are contained in the International Convention on the Prevention of

Pollution from Ships, known as MARPOL 73/78 [121]. In addition, the first sulfur emission-control area (SECA, with a maximum fuel sulfur content of only 1.5% sulfur content) in the Baltic Sea entered into force in 2006, and further SECAs in the North Sea and English Channel entered into force in 2007.

**Fuel Legislation** Engine and vehicle technologies normally achieve their best emissions performance with high-quality fuels [122]. One property on which a great deal of attention has focused is the sulfur content, partly due to the need to reduce PM and SO<sub>2</sub> emissions and partly because fuel sulfur has an adverse effect on certain types of engine and exhaust aftertreatment technology [122]. In Europe, the controls on fuel sulfur content have been periodically tightened, and in the latest step Directive 2003/17/EC required a full transition to “sulfur-free” petrol and diesel (having less than 10 ppm of sulfur) by 2009. This should enable advanced technologies – such as lean-burn engines, particle traps, and regenerative NO<sub>x</sub> storage systems – to meet stringent exhaust emission limits.

Because of the contribution of shipping to pollution at large ports, EU Directive 2005/33/EC limits the fuel sulfur content to 0.1% in marine fuels for harbor regions in 2010.

### Improved Combustion and Exhaust Aftertreatment

It is expected that petrol and diesel will power road transport for the foreseeable future, and considerable reductions in emissions might still be realized in the short term through improved engine design, fuel injection systems, electronic timing, etc. Developments in ship engine technology include direct water injection and humid air motors (in which water vapor is used to cool the charge air and reduce the NO<sub>x</sub> formation during combustion) [96].

The main exhaust aftertreatment technologies that enable vehicle and engine manufacturers to comply with the exhaust emission legislation are three-way catalysts, oxidation catalysts, exhaust gas recirculation (EGR), DPFs, and SCR. Whilst such technologies are generally fitted during manufacture, retrofitting is a common pollution-reduction strategy. For example, DPFs have been fitted to all of London’s 8,000 buses.

In EGR the recirculation of the exhaust air into the combustion chamber reduces the amount of oxygen

available for NO<sub>x</sub> formation. It has been fitted to light-duty diesel road vehicles in Europe for some years. For trains, EGR can give significant reduction in emissions but can be relatively expensive to retrofit. It may also not be possible to fit EGR to many classes of train due to space restrictions [123]. EGR is also applicable to ships but is not yet in regular use.

DPFs physically capture particles in the exhaust stream. They can be broadly divided into two types: “full flow” and “partial flow.” The former lead to a reduction in PM emissions of more than 95%, whereas the latter are more permeable and less effective, but need no maintenance. In order to prevent the filter from blocking the particles captured by the filter must be periodically removed (a process referred to as “regeneration”). The regeneration may be induced using, for example, a heater, fuel burner, or a reactant (such as NO<sub>2</sub>) generated in situ. DPF technologies are also suitable for use on rail and marine engines [124]. However, they can be very bulky and for rail locomotives there are questions concerning the feasibility of retrofitting. Moreover, they appear to be a less cost-effective measure than re-engining [123].

SCR will be the main technology for enabling compliance with future NO<sub>x</sub> emission standards. The method works by injecting ammonia or urea into an exhaust stream to chemically reduce NO<sub>x</sub> to nitrogen. Again, SCR systems are very bulky; the catalyst itself can be considerably larger than the engine. Space is also required for oxidation catalysts and the tank for storing urea or ammonia. This puts severe restrictions on its potential for use as a retrofit item to rail vehicles – where spare space and weight availability is limited – and again re-engining appears to be a more cost-effective option [123]. SCR is in use in a few ships and is still being developed. As well as being an effective NO<sub>x</sub>-reduction measure, one study has also shown that SCR can reduce ship PM emissions by around 50% [35]. No limitations are thought to have been encountered regarding the fitting of SCR to different types of ship, and it can be integrated with other aftertreatment devices [96]. However, sufficiently high temperatures in the exhaust are required for the SCR process to function efficiently. Consequently, large NO<sub>x</sub> reductions are mainly achieved at higher engine loads. Other drawbacks include the space and additional weight requirements, and the potential for “ammonia

slip.” This occurs when excess urea or ammonia is injected into the exhaust stream, and any unused reagent is emitted from the exhaust. The risk of ammonia slip is higher for transient load changes (i.e., maneuvering) [96].

Another shipping technology is seawater scrubbing, in which the exhaust gas is passed through water. This is effective at reducing emissions of SO<sub>2</sub> and probably has benefits for other pollutants.

**Alternative Fuel and Vehicle and Technologies** The development of technology in the longer term will undoubtedly be driven more by the need to reduce CO<sub>2</sub> emissions than by the need to improve air quality. In many cases, a “win-win” outcome is possible, although there can be some conflicts. For example, some studies have indicated that NO<sub>x</sub> emissions can increase as a result of biofuel use [125].

Alternative fuels include liquefied petroleum gas (LPG), compressed natural gas (CNG), and various biofuels (biomethane, ethanol, esterified vegetable oils, etc.). For road transport, limited supplies mean that LPG could never fully replace petrol or diesel, and the scale of uptake of CNG has been impaired by the cost of converting car engines [126]. The availability of suitable engines for rail applications is uncertain; available natural gas engines can also only cover the lower end of the power range needed for railway applications, and there appears to be no use of LPG [123]. Around 50 seagoing ships are currently powered by natural gas, with a reduction in emissions of PM and NO<sub>x</sub> of at least 70% [127]. Biofuels can be used in low concentration blends (5–10%) with petrol or diesel without engine modification, and in high concentrations (85–100%) with specially adapted vehicles. The main benefit of biofuels is their contribution to the reduction of CO<sub>2</sub> emissions; its effects on NO<sub>x</sub> and PM are less clear. Barriers to increased biofuel use include higher costs, compatibility with conventional fuels, and availability of technology for more advanced biofuels [15]. In addition, there have been warnings of hidden costs and adverse impacts on the environment and society, which could arise from the large-scale production of biofuels [128]. It has also been noted that the use of alternative fuels is context-specific; a fuel that is appropriate to one location may not be appropriate to another [23].

Automobile propulsion technologies that are currently under development include “mild” and “full” hybrid vehicles, “plug-in” hybrid vehicles, battery electric vehicles, hydrogen ICE vehicles, and fuel cell vehicles. Vehicle technologies are currently available that can reduce CO<sub>2</sub> emissions per car by 30%, but demand-side and supply-side barriers are delaying their deployment [126]. Cars that emit 50% less CO<sub>2</sub> per kilometer than the equivalent current models could be on the road by 2030, subject to advances in hybrid and battery technologies and industry overcoming cost barriers [126]. In the longer term, a 90% reduction in emissions is possible based on battery-electric propulsion systems [126].

In the rail sector, hybrid systems have been shown to result in reductions in exhaust NO<sub>x</sub> and PM of up to 90% [123]. Hybrid propulsion has been used in shipping for many years, where the mechanical drive for the propellers has been replaced by electrical connections to electric motors, which in turn drive the propellers [129].

Electricity results in zero emissions at the point of use, giving it major local air quality benefits. Moreover, if it is produced from renewable energy it can have low or even effectively zero CO<sub>2</sub> emissions over its life cycle. The main drawbacks with electric road vehicles include a relatively low maximum speed, a short range, the lengthy recharging times and the cost [126]. However, there are continual improvements in the performance and marketability of electric vehicles.

In hydrogen vehicles the power is derived from either combustion or electrochemical conversion in a fuel cell. In hydrogen ICEs the combustion process is fundamentally the same as that in petrol ICEs. In a fuel cell the hydrogen reacts with oxygen to produce water and electricity, and the latter is used to power the vehicle. When hydrogen is used to generate power the only significant emission is water vapor (though a small amount of NO<sub>x</sub> is produced by hydrogen ICEs). If the hydrogen is made from low-CO<sub>2</sub> energy sources, very low life-cycle emissions are possible. Some of the main obstacles to the wider use of hydrogen for transport are the lack of an adequate infrastructure for storage/distribution and the high cost of implementation [15].

**Eco-Driving** “Eco-driving” has been widely publicized as a means of reducing the fuel consumption

and emissions of road vehicles. It is aimed at both private motorists and fleet operators, and typically involves either a simple set of rules to be followed or a program of training. The advice or training varies considerably in terms of the level of detail, but it generally features a number of common actions, including keeping the tires at the correct pressure, reducing the vehicle weight, avoiding sharp acceleration and heavy braking, driving in the highest gear, and avoiding unnecessary engine idling. Average overall reductions in fuel consumption of around 5–10% are typically reported for eco-driving, and some specific examples are shown in Table 2. However, it should be noted that some adverse effects of eco-driving have been observed, such as an increase in NO<sub>x</sub> emissions from diesel cars during urban driving [133].

It is possible that the principles associated with eco-driving on the road are transferable, with modification to the actual actions involved, to rail transport and shipping.

Leaving diesel engines in idle operation is a widespread practice at many rail termini. Reasons for this include keeping the engines warm and providing power brakes, heating/cooling, lights, catering facilities, and

other auxiliaries. Whilst pollutant emissions per train at stations are relatively low compared with those during a trip, where several trains are running at idle and the station is enclosed there may be an issue regarding population exposure. There may therefore be significant potential to reduce emissions at stations simply through enforcing maximum idle running times [123].

When at berth, ships also need electrical power for emergency equipment, refrigeration, cooling, heating, lighting, loading/unloading equipment, etc. This power is normally provided by the auxiliary engines. However, such engines result in air pollutant emissions at the port, and these are often close to centers of population. An alternative approach is to obtain power from shore, a process known as “cold ironing.” The port of Gothenburg in Sweden was the first in the world to provide a high-voltage onshore power supply for cargo vessels. Two local wind turbines are used to generate the electricity, and it is claimed that the system cuts emissions by 94–97% for ships at berth and has led to dramatic reductions in pollutant concentrations. Systems are also available for optimizing the energy efficiency of a vessel by advising the crew on the optimal speed for the prevailing water conditions [138].

**Air Quality, Surface Transportation Impacts on. Table 2**  
Effects of eco-driving measures on fuel consumption

| Action                                        | Change in fuel consumption | References     |
|-----------------------------------------------|----------------------------|----------------|
| Correction of low tire pressure               | –2%                        | [130]          |
| Removal of 100 kg load from a medium-size van | –7%                        | [130]          |
| Removal of roof rack                          | –20%                       | [130]          |
| Not driving with windows open                 | –10%                       | [131]          |
| Not using air-conditioning                    | –18%                       | [131]          |
| Eco-driving gear-change strategy              | –18%                       | [132]          |
| Use of a gear-change indicator                | –1% to –11%                | [133, 134]     |
| Eco-driving training                          | +2% to –16%                | [133, 135–137] |

**Road Sweeping and Washing** Sweeping (or vacuuming) and washing have been applied in several countries as means of decreasing the silt loading of roads, the premise being that this will reduce PM resuspension. Purpose-built vehicles are usually employed, and the cleaning process can involve mechanical (broom) sweeping, vacuuming, spraying with water, or a combination of these. Some vehicles are specifically designed to reduce PM<sub>10</sub> concentrations. A simpler approach is the spraying of water directly onto the road surface (Fig. 14) in an attempt to reduce resuspension on dry days.

Studies in the 1980s showed promising results in terms of the removal of particles from the road surface [139, 140]. However, the majority of later studies have indicated only limited effects on PM<sub>10</sub> concentrations [141–145], and in some cases increases [146]. Where reductions in PM concentrations have been observed, these have generally been small and short-lived (a few hours) [147, 148]. Possible explanations for such limited effectiveness include the silt loading being rapidly



**Air Quality, Surface Transportation Impacts on. Figure 14**

Spraying the road with water in Nijmegen, the Netherlands (Photograph reproduced by kind permission of the Municipality of Nijmegen, the Netherlands)

replaced after sweeping, the frequency of sweeping/washing being too low, and the removal of only particles which are larger than  $PM_{10}$  [145]. The spatial distribution of material on roadways is also highly variable [149].

However, some investigations have indicated effects that are large enough to warrant the consideration of sweeping or washing for reducing roadside  $PM_{10}$  concentrations [150, 151], and if sweeping or washing can remove particles that could evolve into  $PM_{10}$ , then sweeping may have a delayed beneficial effect on air quality [144]. It should be technically undemanding, and relatively low cost. Clearly, both sweeping and washing programs would have to be conducted during periods when traffic flows are low, so that any disruption to the traffic is minimized.

**Dust Suppressants** Dust suppressants (or binders) appear to be an obvious choice for the reduction of resuspension due to road vehicles. Suppressants are chemicals applied to surfaces to maintain the moisture levels or to actually chemically bind the surface material to reduce fugitive dust emissions. The main suppressants in use are water (the oldest, and undoubtedly the cheapest) and solutions of salts such as calcium chloride ( $CaCl_2$ ), magnesium chloride ( $MgCl_2$ ) and

calcium magnesium acetate (CMA). Little quantitative information exists on their effects on airborne particulate matter, although with the need to control  $PM_{10}$  the interest in suppressants is growing. There is reasonably convincing evidence from recent pilot studies in Scandinavia that the use of dust suppressants could make a significant difference to roadside  $PM_{10}$  concentrations on timescales of hours to days [144, 145, 152]. However, more information is needed on potential adverse effects of using dust suppressants on paved roads. A notable concern is road safety, as road surface friction is reduced when large amounts of suppressants are used.

**Traffic Management** Numerous forms of traffic management offer the possibility of reducing emissions from traffic. Local restrictions to the access and/or circulation of traffic have been introduced in many cities to reduce pollutant concentrations. Such restrictions can take a range of forms, including tolls, congestion charges, alternate number plate schemes, and weight restrictions. Cities that have implemented these types of schemes include London, Stockholm, Athens, Budapest, and Prague. Low emission zones (LEZs) are one of the more effective types of restriction. Entry to a LEZ is usually conditional on a vehicle



meeting specified standards. These standards can be set in various different ways, such as emission legislation, the use of specific exhaust aftertreatment (e.g., DPF), or vehicle age.

The potential beneficial effects of such measures can be illustrated by reference to a trial road charging scheme introduced in Stockholm City Centre in 2006. It was estimated that the scheme resulted in a 15% reduction in total road use within the charging zone. Emissions of NO<sub>x</sub> and PM<sub>10</sub> from road traffic in the zone fell by 8.5% and 13%, respectively. Air quality dispersion modelling was applied to assess the effects of the emission reductions on ambient concentrations. It was found that the annual average NO<sub>x</sub> concentrations would decrease by up to 12% along the most densely trafficked streets and PM<sub>10</sub> concentrations would decrease by up to 7% [153].

Other types of traffic management are designed to impose low-emission driving styles, including variable speed limits and synchronized traffic signals (so-called green waves).

### Mitigation Measures

The possibilities for mitigation are less diverse than those for prevention. Mitigation measures typically involve the use of physical barriers, vegetation, or some form of air treatment.

**Physical Barriers** On open roads the emissions from traffic are subject to the normal processes of atmospheric dilution and dispersion. Barriers are common features alongside roads with high levels of traffic, particularly those which run through populated areas. These features are usually introduced to attenuate noise, but can also block pollutant dispersion, increase turbulence and initial mixing, or modify deposition. Any air flow perpendicular to the barrier can be deflected upward by the structure, and this can increase the apparent release height of the pollutant and increase vertical mixing due to the flow separation at the top of the barrier. A recirculation cavity forms in the lee of the structure, and this can also reduce pollutant concentrations relative to an open area with no barrier [155]. However, plume reattachment – and increased pollutant concentrations – can take place further downwind [155].

Inside a road tunnel the dilution and dispersion of pollution is restricted, and unless there is the injection of fresh air, the concentration of pollutants will increase from the entrance to the exit, and the emissions will be concentrated at the portals or ventilation points. Therefore, the road traffic source of pollution is effectively removed along the length of the tunnel and the exterior air quality ought to improve. However, pollutant concentrations can be particularly high in the vicinity of the exit portals. The air quality at the portals or ventilation points will depend upon the level of traffic and whether or not the polluted air is treated in some way before release. When considering the potential for tunnels for improving ambient concentrations, it should be borne in mind that vehicle occupants and tunnel workers may be exposed to increased concentrations, which could be detrimental in public health terms.

**Vegetation** The planting of vegetation can have a positive impact on air quality in the vicinity of roads [156–158]. A number of mechanisms are involved, but the main effects are dry deposition at the leaf surface and the modification of dispersion. The height, width, type, and porosity (leaf density) of the vegetation appear to be important, and greater deposition can occur on coniferous plants than on broadleaf plants due to the larger leaf surface area per unit ground area [159]. However, the results of trials have been mixed, and there are likely to be several practical difficulties associated with the use of vegetative barriers in built-up areas, including available space, soiling by traffic, and the general hostility of the roadside environment to plant life. In addition, analyses carried out for a busy arterial road show that very large vegetation areas (in excess of 10,000 m<sup>2</sup>) would be needed to compensate for the local emissions from vehicles [160].

**Air Treatment** Various air treatment approaches have been devised with the aim of reducing air pollution. In recent years considerable efforts have been made to develop construction materials and coatings, which have the potential for reducing air pollution. These materials are often reliant upon photocatalysis, most commonly involving the compound titanium dioxide (TiO<sub>2</sub>). Titanium dioxide acts as a catalyst to

oxidize NO and NO<sub>2</sub> into nitrate (NO<sub>3</sub><sup>-</sup>) in the presence of oxygen and ultraviolet radiation. The nitrate is subsequently washed away by rainfall. The effectiveness of TiO<sub>2</sub> coatings has been successfully demonstrated in laboratory trials [161], but field trials have been less convincing. This is partly because it is difficult to distinguish the impact of chemical conversion on the catalyst surface from fluctuations in concentrations due to, for example, meteorology.

Several air treatment technologies have been developed for use in tunnels. For example, a number of denitrification systems are available; most work by either a chemical absorption or catalytic process. Tests in Norwegian and Japanese tunnels have indicated that removal rates for NO<sub>2</sub> of up to 90% are possible [162, 163]. The application of denitrification systems at open roadside locations does not appear to have been reported, and appears to be impractical at present. Electrostatic precipitators (ESPs) have been demonstrated as being effective in removing particles from tunnel air with varying efficiency. Some electrostatic systems have also been developed for exterior environments and tunnels [164], but further investigation into their effectiveness is required.

## Future Directions

### Outlook

The future global outlook for transport-related pollution currently looks rather bleak. In developing countries, the rise in car ownership and travel is predicted to be even more dramatic than it has been in industrialized nations. For example, the rate of car ownership in Brazil is predicted to overtake that in the country that has historically been the world leader – the United States – before the middle of the century. Car ownership in China has doubled in the last 5 years, and it already has the third highest car sales in the world [126].

However, at the time of writing there appear to be some signs of change, but probably not as a result of environmental pressures. Certain Member States of the EU reported rather large decreases in emissions from road transport between 2007 and 2008, reflecting at least in part the lower amount of freight transport by road in the second half of 2008 as a result of the economic recession [8]. In addition, the

recession has had a significant effect on car sales. Several governments have introduced scrappage schemes to maintain or stimulate demand for new cars, but these schemes should also result in the replacement of older polluting vehicles with new and more efficient ones. Data sets from other regions also show the impact of the economic climate. In 2008, in the United States, the number of vehicles per 1,000 people decreased for the first time since 1997. Similarly, total vehicle-miles travelled decreased slightly from 2007 to 2008 – the first decline since 1980 [165].

Nevertheless, governments around the world are acknowledging that action is required to control the rate of traffic growth, to improve the environmental performance of vehicles, and to increase public awareness about the impacts of pollutant emissions from transport. As noted earlier, it is likely that the future shape of transport will be influenced more by the need to reduce greenhouse gas emissions than the need to improve local and regional air quality. The development of a sustainable transport system will involve a large number of interventions that are beyond the scope of this entry, but will probably include reducing dependence on cars, effective land-use planning to reduce journey distances, improving the quality of public transport, etc. Preventative measures will also need to be accompanied by mitigation measures which reduce human exposure to air pollution.

For the control of local air pollution, it is likely that emission legislation will still be the main instrument for governments in the future. However, the future of European emission legislation post-Euro 6/VI is somewhat uncertain. There is a drive toward the global harmonization of legislation, but given that the threat of climate change is increasing the pressure on manufacturers to develop low-carbon power sources as alternatives to fossil fuels (batteries, fuel cells, etc.), it is therefore unlikely that the current test procedures will remain relevant.

Shipping is currently one of the least regulated emission sources, with a high reduction potential through technological improvements, alternative fuels, and ship modifications. It has been observed that if the sulfur content of fuel oil remains at present-day levels, a doubling in the emissions of some pollutants from shipping can be expected by 2050. However, given the increasing concern about shipping-related air quality,

further reductions of NO<sub>x</sub> and SO<sub>2</sub> emissions seem likely. Using aggressive NO<sub>x</sub> emission reduction technologies, a decrease up to 85% in NO<sub>x</sub> emissions could be reached through technological improvements by 2050, in spite of a growing fleet [166].

### Research Needs

Emissions from transport, and the resulting ambient air pollution, have been studied intensively, although there are still a number of areas where the understanding is poor or incomplete. Some of these areas are summarized below.

**Testing Methods for New Technologies** In legislative terms, the emphasis at present is on the definition of appropriate procedures for the type approval of hybrid-electric, full-electric vehicles and hydrogen ICE vehicles. In some cases, certification will be problematic. For example, the current heavy-duty engine test procedure cannot be applied meaningfully to hybrid vehicles, as the instantaneous behavior of the engine is decoupled from the instantaneous road load. Vehicle-based tests may therefore become the norm in the future, although this in itself introduces certain challenges given the large number of possible vehicle configurations and differences in real-world operation. Such challenges might be overcome by the modelling of powertrains.

**Understanding PM Emissions** In recent years evidence has accumulated indicating that airborne particles exert a range of adverse health effects. These effects are diverse in scope, severity, and duration [4]. More work is needed in the future to accurately characterize particle emissions from surface transport modes and to uncover the mechanisms by which health impacts occur.

It has already been noted in this entry that a better understanding of non-exhaust PM is required, including characterization of the emitted particles, investigation of health impacts, and the development of methods of prevention and mitigation. The importance of brake wear may diminish in the future as regenerative braking systems in hybrid and electric vehicles partially or fully replace traditional mechanical ones. However, this is not currently taken into account in emission calculations and projections.

More detailed information on PM emissions from rail and shipping (especially inland shipping) is also required, including emission factors for different engines and fuel type combinations, in terms of size distribution, particle numbers, and chemical speciation [34]. Few studies have examined PM levels in the ambient air alongside railway lines. These often run very close to housing, and may contribute to local air quality problems, especially where large numbers of diesel trains are in use. PM concentrations could also be elevated near shunting yards and depots. In addition, there is little information on PM concentrations in stations where diesel trains are left running at idle and the dispersion of pollution is restricted.

**Understanding NO<sub>x</sub> Emissions** In order to understand atmospheric NO<sub>2</sub> it is important to understand NO<sub>x</sub> emissions, and it is possible that the latter are being underestimated by models. For example, there is evidence to suggest that NO<sub>x</sub> emissions from Euro V heavy-duty vehicles are significantly higher than the limit at type approval [110]. Vehicles equipped with SCR have very low emission levels during rural and motorway driving, but not under the low engine loads associated with urban or driving. The reason for this is that the SCR system needs to attain a certain temperature for efficient NO<sub>x</sub> reduction. The real-world performance of SCR should therefore be investigated. Furthermore, more extensive information on f-NO<sub>2</sub> values for a wider range of vehicles is required.

**Improving Emission Models** Accuracy, reliability, consistency, and credibility are prerequisites of emission estimates. However, comparisons between the results from different emission models have highlighted substantial differences. There is therefore a need for emission modelling approaches to be reevaluated, and for improvements to be made. This requires a better understanding of the circumstances under which emission models provide inaccurate predictions. For example, some modellers believe that the exhaust emission factors for road vehicles are unsuitable for certain road types and traffic conditions. Particular concerns have been expressed about the validity of the emission factors for so-called congested traffic. The term “congestion” also needs to be defined in a manner that is meaningful in terms of emissions.

**Improving Activity Data** While obtaining primary activity data for road transport by direct measurement remains the main approach for short-term modelling exercises, it tends to be a rather expensive activity, and would be unfeasible where an emission estimate is required for a large geographical area. Traffic assignment models and microsimulation tools are increasingly being used to provide the relevant data, and communications technology such as Intelligent Transport Systems and Advanced Transport Telematics enable transport networks to be monitored in real time. It could well be possible to adapt a number of models and new technologies to contribute appropriate information on fleets and traffic activity.

In shipping the use of data from AIS and CEMS has a great deal of potential, and even more so if it can also be extended to cover inland shipping. Data from AIS have already been incorporated into at least one model [167]. However, some work is required to validate emission estimates obtained using AIS data [34]. Research is also required on the movements of empty ships, as these are often excluded from emission estimates [34].

**Model Validation** Model validation is an important area for improvement. The literature on the validation of transport emission models is quite limited, and validation studies are often only partial in the sense that they address just one type of road, type of vehicle, or time period. The understanding of the uncertainties in emission models and the main factors affecting prediction errors is therefore poor [168]. Emission models should quantify prediction errors. The models in common use should also be compared for a wide range of traffic and fleet conditions and locations. This would allow for a better assessment of the relative performance of different model types.

**Evaluation of Mitigation Measures** The results of field trials on pollution mitigation methods have been rather mixed, and are probably confounded by site-specific factors. Consequently, much more work is required to determine which measures are likely to be most effective for a given set of circumstances.

## Bibliography

### Primary Literature

1. Bell ML, Davis DL, Fletcher T (2004) A retrospective assessment of mortality from the London smog episode of 1952: the role of influenza and pollution. *Environ Health Perspect* 112(1):6–8
2. Haagen-Smit AJ (1952) Chemistry and physiology of Los Angeles smog. *Ind Eng Chem* 44(6):1342–1346
3. COMEAP (2009) Long-term exposure to nitrogen dioxide: epidemiological evidence of effects on respiratory morbidity in children. Committee on the Medical Effects of Air Pollutants. Subgroup: Quantification of Air Pollution Risks II
4. Morawska L, Moore MR, Ristovski ZD (2004) Health impacts of ultrafine particles – desktop literature review and analysis. Australian Department of the Environment and Heritage, Canberra. ISBN 0642550557
5. Krasenbrink K, Martini G, Wass U, Jobson E, Borken J, Kuehne R, Ntziachristos L, Samaras Z, Keuken M (2005) Factors determining emissions in the WHO European Region. In: Krzyzanowski M et al (eds) Health effects of transport-related air pollution. Regional Office for Europe of the World Health Organization, Copenhagen. ISBN 92 890 1373 7
6. Corbett JJ, Wang C, Winebrake JJ, Green E (2007) Allocation and forecasting of global ship emissions. Submitted as a formal annex submission to the International Maritime Organization. Clean Air Task Force and Friends of the Earth International, Boston, 11 Jan 2007, pp 26
7. Eyring V, Köhler HW, van Aardenne J, Lauer A (2005) Emissions from international shipping: 1. The last 50 years. *J Geophys Res* 110(D17):D17305
8. EEA (2010) European Union emission inventory report 1990–2008 under the UNECE Convention on Long-range Transboundary Air Pollution (LRTAP). EEA Technical Report No 7/2010. European Environment Agency, Copenhagen
9. AIRPARIF (2001) Etude de la qualité de l'air dans le secteur de la gare de l'Est. Paris, AIRPARIF – Surveillance de la Qualité de l'Air en Ile-de-France, Laboratoire d'Hygiène de la Ville de Paris, Mairie de Paris, DASES, Laboratoire Central de la Préfecture de Police
10. Seaton A, Cherrie J, Dennekamp M, Donaldson K, Hurley JF, Tran CL (2005) The London underground: dust and hazards to health. *Occup Environ Med* 62(6):355–362
11. Branis M (2006) The contribution of ambient sources to particulate pollution in spaces and trains of the Prague underground transport system. *Atmos Environ* 40:348–356
12. Raut J-C, Chazette P, Fortain A (2009) Link between aerosol optical, microphysical and chemical measurements in an underground railway station in Paris. *Atmos Environ* 43:860–868
13. Corbett JJ, Winebrake JJ, Green EH, Kasibhatla P, Eyring V, Lauer A (2007) Mortality from ship emissions: a global assessment. *Environ Sci Technol* 41:8512–8518
14. Hulscombe JHJ, Denier van der Gon HAC (2010) Fuel consumption and associated emissions from seagoing ships at

- berth derived from an on-board survey. *Atmos Environ* 44:1229–1236
15. Uherek E et al (2010) Transport impacts on atmosphere and climate: land transport. *Atmos Environ* 44(37):4772–4816
  16. European Commission (2003) European Union energy and transport in figures – statistical pocketbook 2003. European Commission, Brussels
  17. European Commission (2010) European Union energy and transport in figures – statistical pocketbook 2010. European Commission, Brussels
  18. Keuken M, Sanderson E, van Aalst R, Borken J, Schneider J (2005) Contribution of traffic to levels of ambient air pollution in Europe. In: Krzyzanowski M et al (eds) Health effects of transport-related air pollution. Regional Office for Europe of the World Health Organization, Copenhagen. ISBN 92 890 1373 7
  19. EEA (2007) Air pollution in Europe 1990–2004. EEA Report No 2/2007. European Environment Agency, Copenhagen
  20. AQEG (2004) Nitrogen dioxide in the United Kingdom. Air Quality Expert Group. Published by the Department for Environment, Food and Rural Affairs. DEFRA Publications, London
  21. Harrison RM, Stedman J, Derwent D (2008) Why are PM<sub>10</sub> concentrations in Europe not falling? New directions, atmospheric science perspectives special series. *Atmos Environ* 42:603–606
  22. Minoura H, Ito A (2010) Observation of the primary NO<sub>2</sub> and NO oxidation near the trunk road in Tokyo. *Atmos Environ* 44:23–29
  23. Gorham R (2002) Air pollution from ground transportation an assessment of causes, strategies and tactics, and proposed actions for the international community. Department of Economic and Social Affairs, United Nations
  24. Duan N (1982) Models for human exposure to air pollution. *Environ Int* 8:305–309
  25. Ott WR (1982) Concepts of human exposure to air pollution. *Environ Int* 7:179–196
  26. Heywood JB (1988) Internal combustion engine fundamentals. McGraw-Hill, London
  27. Vestreng V, Ntziachristos L, Semb A, Reis S, Isaksen ISA, Tarrasón L (2009) Evolution of NO<sub>x</sub> emissions in Europe with focus on road transport control measures. *Atmos Chem Phys* 9:1503–1520
  28. Kittelson DB (1998) Engines and nanoparticles: a review. *J Aerosol Sci* 29:575–588
  29. Abdul-Khalek IS, Kittelson DB, Brear F (1999) The influence of dilution conditions on diesel exhaust particle size distribution measurements. SAE Technical Paper Series No. 1999-01-1142
  30. Mathis U (2002) Influencing parameters of nanoparticle formation from diesel exhaust. In: Proceedings of the 6th international ETH conference on nanoparticle measurements, Zürich, 19–21 August 2002
  31. Ntziachristos L, Samaras Z, Pistikopoulos P, Kyriakis N (2000) Statistical analysis of diesel fuel effects on particle number and mass emissions. *Environ Sci Technol* 34:5106–5114
  32. Maricq MM, Chase RE, Podsiadlik DH (1999) Vehicle exhaust particle size distributions: a comparison of tailpipe and dilution tunnel measurements. SAE Technical Paper Series No. 1999-01-1461
  33. EEA (2009) The EMEP/EEA air pollutant emission inventory guidebook – 2009. EEA Technical Report No 9/2009. European Environment Agency, Copenhagen
  34. Denier van der Gon H, Hulskotte J (2009) Methodologies for estimating shipping emissions in the Netherlands – a documentation of currently used emission factors and related activity data. Report 500099012, Netherlands Environmental Assessment Agency, Bilthoven. ISSN: 1875-2322
  35. Fridell E, Steen E, Peterson K (2008) Primary particles in ship emissions. *Atmos Environ* 42:1160–1168
  36. CONCAWE (1987) An investigation into evaporative hydrocarbon emissions from European vehicles. Report No. 87/60. CONCAWE, The Hague
  37. Thorpe A, Harriosn RM (2008) Sources and properties of non-exhaust particulate matter from road traffic: A review. *Sci Total Environ* 400:270–282
  38. Pierson WR, Brachaczek WW (1974) Airborne particulate debris from rubber tyres. *Rubber Chem Technol* 47:1275–1299
  39. Rogge WF, Hildemann LM, Mazurek MA, Cass GR (1993) Sources of fine organic aerosol. 3. Road dust, tire debris, and organometallic brake lining dust: roads as sources and sinks. *Environ Sci Technol* 27(9):1892–1904
  40. Fauser P (1999) Particulate air pollution with emphasis on traffic generated aerosols. Riso-R-1053(EN), Riso National Laboratory, Roskilde
  41. Cardina JA (1974) Particle size determination of tire tread rubber in atmospheric dusts. *Rubber Chem Technol* 47:1005–1010
  42. Dannis ML (1974) Rubber dust from the normal wear of tyres. *Rubber Chem Technol* 47:1011–1037
  43. Cadle SH, Williams RL (1978) Gas and particle emissions from automobile tires in laboratory and field studies. *Rubber Chem Technol* 52:146–158
  44. Rauterberg-Wulff A (1998) Beitrag des Reifen- und Bremsenabriebs zur Russimmission an Strassen. Dissertation am Fachgebiet Luftreinhaltung der Technischen Universität Berlin
  45. Rauterberg-Wulff A (1999) Determination of emission factors for tyre wear particles up to 10 µm by tunnel measurements. In: Proceedings of 8th international symposium “Transport and Air Pollution”, Graz, 31 May–2 June 1999. Technical University Graz. Institute for Internal Combustion Engines and Thermodynamics
  46. Moosmüller H, Gillies J, Rogers F, DuBois D, Chow J, Watson J, Langston R (1998) Particulate emission rates for unpaved shoulders along a paved road. *J Air Waste Manage Assoc* 48:398–407
  47. Dahl A, Gharibi A, Swietlicki E, Gudmundsson A, Bohgard M, Ljungman A, Blomqvist G, Gustafsson M (2006) Traffic-generated emissions of ultrafine particles from pavement-tire interface. *Atmos Environ* 40:1314–1323
  48. Berdowski JJM, Mulder W, Veldt C, Visschedijk AJH, Zandveld PYJ (1997) Particulate matter emissions (PM<sub>10</sub> – PM<sub>2.5</sub> – PM<sub>0.1</sub>) in Europe in 1990 and 1993. TNO Report TNO\_MEP – R 96/472.



- TNO Institute of Environmental Sciences, Energy Research and Process Innovation, Apeldoorn
49. Sanders PG, Xu N, Dalka TM, Maricq MM (2003) Airborne brake wear debris: size distributions, composition, and a comparison of dynamometer and vehicle tests. *Environ Sci Technol* 37:4060–4069
  50. Garg BD, Cadle SH, Mulawa PA, Groblicki PJ, Laroo C, Parr GA (2000) Brake wear particulate matter emissions. *Environ Sci Technol* 34(21):4463–4469
  51. Abu-Allaban M, Gillies JA, Gertler AW, Clayton R, Proffitt D (2003) Tailpipe, resuspended road dust, and brake wear emission factors from on-road vehicles. *Atmos Environ* 37(1):5283–5293
  52. Iijima A, Sato K, Yano K, Tago H, Kato M, Kimura H, Furuta N (2007) Particle size and composition distribution analysis of automotive brake abrasion dusts for the evaluation of anti-monony sources of airborne particulate matter. *Atmos Environ* 41:4908–4919
  53. Bukowiecki N, Gehrig R, Hill M, Lienemann P, Zwicky CN, Buchmann B, Weingartner E, Baltensperger U (2007) Iron, manganese and copper emitted by cargo and passenger trains in Zurich (Switzerland): Size-segregated mass concentrations in ambient air. *Atmos Environ* 41:878–889
  54. Gehrig R, Hill M, Lienemann P, Zwicky CN, Bukowiecki N, Weingartner E, Baltensperger U, Buchmann B (2007) Contribution of railway traffic to local PM<sub>10</sub> concentrations in Switzerland. *Atmos Environ* 41:923–933
  55. Kleeman MJ, Cass GR (1999) Effect of emissions control strategies on the size and composition distribution of urban particulate air pollution. *Environ Sci Technol* 33:177–189
  56. Gaffney P, Bode R, Murchison L (1995) PM<sub>10</sub> emission inventory improvement program for California. Air Resources Board, Sacramento
  57. Zimmer RA, Reeser WK, Cummins P (1992) Evaluation of PM<sub>10</sub> emission factors for paved streets. In: Chow JC, Ono DM (eds) PM<sub>10</sub> standards and non-traditional particulate source controls., pp 311–323
  58. Pregger T, Friedrich R (2002) Emission Sources of PM<sub>10</sub> (coarse and fine fraction) in Germany, Universität Stuttgart, Stuttgart. Presentation at "Particles in the size of 2.5 to 10 microns in urban areas" workshop, in support of the Clean Air for Europe (CAFE) programme of DG ENV, Berlin, 4–6 Nov 2002
  59. Düring I, Lohmeyer A (2003) Quantifizierung der PM<sub>10</sub>-Emissionen durch Staubaufwirbelung und brieb von Straßen auf der Basis von vorhandenen Messdaten. <http://www.lohmeyer.de>
  60. Thorpe AJ, Harrison RM, Boulter PG, McCrae IS (2007) Estimation of particle resuspension source strength on a major London Road. *Atmos Environ* 41:8007–8020
  61. Johansson C (2002) Source contributions of PM in Sweden – implications for abatement strategies. Institute of Applied Environmental Research (ITM). Stockholm. Presentation at "Particles in the size of 2.5 to 10 microns in urban areas" workshop, in support of the Clean Air for Europe (CAFE) programme of DG ENV, Berlin, 4–6 Nov 2002
  62. Areskoug H, Johansson C, Alesand T, Hedberg E, Ekengena T, Vesely V, Wideqvist U, Hansson H-C (2004) Concentrations and sources of PM<sub>10</sub> and PM<sub>2.5</sub> in Sweden. ITM Report 110. Institute of Applied Environmental Research, Stockholm University. ISSN 1103-341X
  63. Gustafsson M, Blomqvist G, Gudmundsson A, Dahl A, Johnsson P, Swietlicki E (2009) Factors influencing PM<sub>10</sub> emissions from road pavement wear. *Atmos Environ* 43:4699–4702
  64. Sturm PJ, Boulter P, de Haan P, Joumard R, Hausberger S, Hickman AJ, Keller M, Niederle W, Ntziachristos L, Reiter C, Samaras Z, Schinagl G, Schweizer T, Pischinger R (1998) Instantaneous emission data and their use in estimating passenger car emissions. Deliverable no. 6. Of the EC MEET Project. Technical University of Graz, Institute for Internal Combustion Engines and Thermodynamics, Austria
  65. Van de Weijer C (1997) Heavy duty emission factors. Development of representative driving cycles and prediction of emissions in real life. Ph.D. thesis, Graz University of Technology, Oct 1997
  66. INFRAS (2010) Handbook of emission factors for road transport, version 3.1. INFRAS, Zürich. <http://www.hbefa.net/e/index.html>
  67. Smit R, Smokers R, Schoen E, Hensema A (2006) A new modelling approach for road traffic emissions: VERSIT+LD: background and methodology. Report 06.OR.PT.016.1/RS. TNO Science and Industry, Delft
  68. Hassounah MI, Miller EJ (1995) Modelling air pollution from road traffic: a review. *Traffic Eng Contr* 35(9):510–514
  69. Hung W-T, Tomg H-Y, Cheung C-S (2005) A modal approach to vehicular emissions and fuel consumption model development. *J Air Waste Manage Assoc* 55:1431–1440
  70. Coelho MC, Fariasa TL, Roupail NM (2005) A methodology for modelling and measuring traffic and emission performance of speed control traffic signals. *Atmos Environ* 39:2367–2376
  71. De Haan P, Keller M (2000) Emission factors for passenger cars: application of instantaneous emission modelling. *Atmos Environ* 34:4629–4638
  72. Sorenson SC, Schramm J (1992) Individual and public transportation-emissions and energy consumption models. Technical University of Denmark report, RE 91-5, Lyngby
  73. Hansen JQ, Winter M, Sorenson SC (1995) The influence of driving patterns on petrol passenger car emissions. *Sci Total Environ* 169:129–139
  74. Joumard R, Jost P, Hickman J (1995) Influence of instantaneous speed and acceleration on hot passenger car emissions and fuel consumption. SAE Paper 950928. Society of Automotive Engineers, Warrendale
  75. Post K, Kent JH, Tomlin J, Carruthers N (1984) Fuel consumption and emission modelling by power demand and a comparison with other models. *Transp Res* 18A:191–213
  76. Rexeis M, Hausberger S, Riemersma I, Tartakovsky L, Zvirin Y, Erwin C (2005) Heavy-duty vehicle emissions. Final report of WP 400 in ARTEMIS (Assessment and Reliability of Transport Emission Models and Inventory Systems); DGTREN Contract 1999-RD.10429; University of Technology, Graz, Report no.: I 02/2005/Hb 20/2000 I680

77. Atjay D, Weilenmann M (2004) Compensation of the exhaust gas transport dynamics for accurate instantaneous emission measurement. *Environ Sci Technol* 38(19):5141–5148
78. Roujol S, Jourmard R (2009) Influence of passenger car auxiliaries on pollutant emission factors within the Artemis model. *Atmos Environ* 43(5):1008–1014
79. European Commission (1999) MEET – methodology for calculating transport emissions and energy consumption. Office for Official Publications of the European Communities, Luxembourg
80. André J-M, Jourmard R (2005) Modelling of cold start excess emissions for passenger cars. INRETS Report LTE0509. Laboratoire Transports et Environnement, INRETS, Bron
81. Blaikley DCW, Smith AP, Feest EA, Reading AH (2001) UG219 TRAMAQ – cold-start emissions. Summary report, AEAT/ENV/R/0638 Issue 1, May 2001
82. Wade DT (1967) Factors influencing vehicle evaporative emissions. Esso Research and Engineering, Society of Automotive Engineers paper SAE 670126
83. Hausberger S, Wiesmayr J, Bukvarevic E, Tripold W, Brenner J (2005) Evaporative emissions of vehicles – final report. European Commission 5th Framework project ARTEMIS (Assessment and Reliability of Transport Emission Models and Inventory Systems) Technical University of Graz, Graz
84. Martini G et al (2007) Joint EUCAR/JRC/CONCAWE programme on effects of petrol vapour pressure and ethanol content on evaporative emissions from modern cars. Report number EUR 22715 EN. ISBN 978-1-4419-0851-3\_554. Office for Official Publications of the European Communities, Luxembourg
85. Mellios G, Samaras Z (2007) An empirical model for estimating evaporative hydrocarbon emissions from canister-equipped vehicles. *Fuel* 86:2254–2261
86. USEPA (1995) Compilation of air pollutant emission factors AP-42, 5th edn. Stationary point and area sources, vol 1. Section 13.2.1: paved roads (updated November 2006). United States Environmental Protection Agency, Washington, DC
87. Venkatram A (2000) A critique of empirical emission factor models: a case study of the AP-42 model for estimating PM<sub>10</sub> emissions from paved roads. *Atmos Environ* 32:1–11
88. Jørgensen MV, Sorenson SC (1996) Estimating emissions from railway traffic. Report for MEET, 4th framework project, (DG-VII contract no: ST-96-SC.204), as Report ET-EO-97-03, Department of Energy Engineering, Technical University of Denmark
89. Georgakaki A, Coffey R, Sorenson SC (2002) Development of a database system for the calculation of indicators of environmental pressure caused by railway transport. Transport and ENVIRONMENT DATABASE SYSTEM (TRENDS) project for EUROSTAT, Report MEK-ET-2002-03, Department of Energy Engineering, Technical University of Denmark
90. Lindgreen E, Sorenson SC (2005) Simulation of energy consumption and emissions from rail traffic. Report MEK-ET-2005-4, Department of Mechanical Engineering, Danish Technical University, February 2005. Deliverable 7a for ARTEMIS, DG-TREN contract 1999-RD.10429
91. Gijzen A, van den Brink RMM (2002) Energy use and emissions by rail transport; description and results of the model PRORIN (in Dutch). Report no. 773002021. RIVM, Bilthoven
92. Entec (2002) Quantification of emissions from ships associated with ship movements between ports in the European Community. Final Report to the European Commission, Entec UK Limited
93. Sjöbris A, Gustafsson J, Jivén K (2005) ARTEMIS: sea transport emission modelling. MariTerm AB, Gothenburg
94. Georgakaki A, Sorensen SC (2004) Report on collected data and resulting methodology for inland shipping. Workpackage 900: emission estimating methodology for ship transport of the European Commission fifth framework ARTEMIS project. Report No: MEK-ET-2004-2. ISBN 87-7475-314-2. Technical University of Denmark, Lyngby
95. André M, Rapone M, Adra N, Pollak I, Keller M, McCrae I (2006) Traffic characteristics for the estimation of pollutant emissions from road transport. Final Report of ARTEMIS Workpackage 1000. INRETS Report LTE0606. INRETS, Bron
96. Stathopoulos FG, Noland RB (2003) Induced travel and emissions from traffic flow improvement projects. *TRR J Transp Res Board* 1842:57–63
97. Schall V (2007) Energy-efficient driving. Presentation at the International Transport Forum workshop on eco-driving, Paris. 22 November 2007. Deutsche Bahn AG
98. Entec (2005) Ship emissions: assignment, abatement and market-based instruments – task 1: preliminary assignment of ship emissions to European countries. Report to the European Commission, Entec UK Limited
99. AQEG (2005) Particulate matter in the United Kingdom. Report of the Air Quality Expert Group, DEFRA, London
100. Matthijsen J, ten Brink HM (2007) PM<sub>2.5</sub> in the Netherlands – consequences of the new European air quality standards. MNP Report 500099001/2007. The Netherlands Environmental Assessment Agency (MNP), Bilthoven
101. Putaud J-P et al (2010) A European aerosol phenomenology – 3: physical and chemical characteristics of particulate matter from 60 rural, urban, and kerbside sites across Europe. *Atmos Environ* 44:1308–1320
102. Harrison RM, Shi JP, Grenfell JL (1998) Novel night-time free radical chemistry in severe nitrogen dioxide pollution episodes. *Atmos Environ* 32(16):2769–2774
103. AQEG (2007) Trends in primary nitrogen dioxide in the UK. The fourth report of the Air Quality Expert Group. Department for the Environment, Food and Rural Affairs, London, ISBN:978-0-85521-179-0
104. Jenkin ME (2004) Analysis of sources and partitioning of oxidant in the UK – part 2: contributions of nitrogen dioxide emissions and background ozone at a kerbside location in London. *Atmos Environ* 38(30): 5131–5138
105. Carslaw DC, Beevers SD (2004) Investigating the potential importance of primary NO<sub>2</sub> emissions in a street canyon. *Atmos Environ* 38:3585–3594

106. Carslaw DC (2005) Evidence of an increasing NO<sub>2</sub>/NO<sub>x</sub> emissions ratio from road traffic emissions. *Atmos Environ* 39:4793–4802
107. Hueglin C, Buchmann B, Weber RO (2006) Long-term observation of real-world road traffic emission factors on a motorway in Switzerland. *Atmos Environ* 40(20): 3696–3709
108. Grice S, Stedman J, Kent A, Hobson M, Norris J, Abbott J, Cooke S (2009) Recent trends and projections of primary NO<sub>2</sub> emissions in Europe. *Atmos Environ* 43:2154–2167
109. Lappi M, Laurikko J, Erkkilä K, Pellikka A-P, Koskentalo T (2008) Specific NO and NO<sub>2</sub> emissions from a wide range of current and future LD and HD vehicles in urban driving conditions. In: *Proceedings of 16th international conference "Transport and Air Pollution" 2008, Graz*
110. Rexeis R, Hausberger S (2009) Trend of vehicle emission levels until 2020 – prognosis based on current vehicle measurements and future emission legislation. *Atmos Environ* 43:4689–4698
111. Keuken M, Roemer M, van den Elshout S (2009) Trend analysis of urban NO<sub>2</sub> concentrations and the importance of direct NO<sub>2</sub> emissions versus ozone/NO<sub>x</sub> equilibrium. *Atmos Environ* 43:4780–4783
112. WHO (2003) Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. Report on a WHO working group, Bonn, 13–15 Jan 2003. Document EUR/03/5042688. WHO Regional Office for Europe, Copenhagen
113. Granier C, Brasseur GP (2003) The impact of road traffic on global tropospheric ozone. *Geophys Res Lett* 30(2):1086. doi:10.1029/2002GL015972
114. Matthes S, Grewe V, Sausen R, Roelofs G-J (2007) Global impact of road traffic emissions on tropospheric ozone. *Atmos Chem Phys* 7:1707–1718, <http://www.atmos-chem-phys.net/7/1707/2007/>
115. Berkowicz R (2000) OSPM – a parameterised street pollution model. *Environ Monit Assess* 65(1/2):323–331
116. GLA (2010) Clearing the air: the mayor's draft air quality strategy for public consultation. Greater London Authority, London
117. Eyring V et al (2009) Transport impacts on atmosphere and climate: shipping. *Atmos Environ* 44(37):4735–4771
118. Harrison RM, Yin J, Mark D, Stedman J, Appleby RS, Booker J, Moorcroft S (2001) Studies of the coarse particle (2.5 – 10 µm) component in UK urban atmospheres. *Atmos Environ* 35:3667–3679
119. Delphi (2010) World-wide emission standards: passenger cars and light duty vehicles (2010/11). Delphi, Troy
120. UIC (2006) Exhaust emission tests for diesel traction engines, Leaflet 624, 3rd edn. Union Internationale des Chemins de Fer, Paris. ISBN 2-7461-1040-7
121. IMO (1998) Regulations for the prevention of air pollution from ships and NO<sub>x</sub> technical code, ANNEX VI of MARPOL 73/78. International Maritime Organization, London
122. ACEA, Alliance, EMA, JAMA (2006) Worldwide fuel charter, 4th edn. September 2006
123. Hill N et al (2005) Rail diesel study. Work package 2 final report: technical and operational measures to improve the emissions performance of diesel rail. Report ED05010/WP2/R02. AEA Technology, Didcot
124. MECA (2006) Case studies of the use of exhaust emission controls on locomotives and large marine diesel engines. Manufacturers of Emission Controls Association, Washington, DC
125. Karavalakis G, Stournas S, Bakeasb E (2009) Effects of diesel/biodiesel blends on regulated and unregulated pollutants from a passenger vehicle operated over the European and the Athens driving cycles. *Atmos Environ* 43(10):1745–1752
126. King J (2007) The King review of low-carbon cars. Part I: the potential for CO<sub>2</sub> reduction. HMSO, London
127. Bruckner-Menchelli N (2009) Scrubbers versus distillates. *Bunkerworld*, May/June 2009, p 24
128. Palmer MW, Pimentel D, Lal R, Kennedy D (2007) Biofuels and the environment. *Science* 317:897–898
129. Katsoulakos T, Morgan D (2009) Ship technologies. Deliverable D2.4.1 of the EC 7th framework programme coordination action SKEMA (Sustainable Knowledge Platform for the European Maritime and Logistics Industry)
130. SenterNovem (2005) Ecodriving – the smart driving style. Produced by SenterNovem, Utrecht for the EC TREATISE project, Sept 2005
131. Van Mieghem R, Gense R (2005) Extra brandstofverbruik door gebruik van accessoires in personenauto's. TNO report 05.OR.VM.066.1/RvM. TNO Science and Industry, Delft
132. Weilenmann M (2002) Emissions and fuel consumption with eco drive – English summary. Report number 201209e. EMPA, Dübendorf
133. Vermeulen RJ (2006) The effects of a range of measures to reduce the tail pipe emissions and/or the fuel consumption of modern passenger cars on petrol and diesel. TNO report IS-RPT-033-DTS-2006-01695. TNO Science and Industry, Delft
134. Fontaras G, Pistikopoulos P, Samaras Z (2008) Experimental evaluation of hybrid vehicle fuel economy and pollutant emissions over real-world simulation driving cycles. *Atmos Environ* 42:4023–4035
135. Johansson H, Gustafsson P, Henke M, Rosengren M (2003) Impact of EcoDriving on emissions. In: *Transport and air pollution. Proceedings from the 12th symposium, Avignon, 16–18 June 2003*
136. Parkes AM, Reed N (2005) Fuel efficiency training in a full-mission truck simulator. In: *Behavioural research in road safety 2005: fifteenth seminar, London, pp 135–146*
137. Zarkadoulas M, Zoidis G, Tritopoulou E (2007) Training urban bus drivers to promote smart driving: a note on a Greek eco-driving pilot program. *Transp Res D Transp Environ* 12(6):449–451
138. Schweighofer J, Seiwert P (2007) Environmental performance of inland navigation. European Inland waterway navigation conference, Visegrád, 27–29 June 2007
139. Duncan M, Jain R, Yung SC, Patterson R (1985) Performance evaluation of an improved street sweeper. *US Environmental*

- Protection Agency (US EPA-600/7-85-008), Government Printing Office, Research Triangle Park, NC
140. Cowherd C, Muleski GE, Kinsey JS (1988) Control of open fugitive dust sources. EPA-450/3-88-008, US Environmental Protection Agency, Research Triangle Park, NC
  141. Chow JC, Watson JG, Egami RT, Frazier CA, Lu Z, Goodrich A, Bird A (1990) Evaluation of regenerative-air vacuum street sweeping on geological contributions to PM<sub>10</sub>. *J Air Waste Manage Assoc* 40(8):1134–1142
  142. Kuhns H, Etyemezian V, Green M, Hendrickson K, McGown M, Bartond K, Pitchford M (2003) Vehicle-based road dust emission measurement – part II: effect of precipitation, wintertime road sanding, and street sweepers on inferred PM<sub>10</sub> emission potentials from paved and unpaved roads. *Atmos Environ* 37:4573–4582
  143. Gustafsson M (2004) Nordic road and transport research, no.1. Road and street sweeping as a measure against high particle concentrations caused by road dust. <http://www.vti.se/nordic/1-04mapp/annosv.htm>
  144. Norman M, Johansson C (2006) Studies of some measures to reduce road dust emissions from paved roads in Scandinavia. *Atmos Environ* 40(32):6154–6164
  145. Aldrin M, Haff I, Rosland P (2007) The effect of salting with magnesium chloride on the concentration of particulate matter in a road tunnel. Norwegian Public Road Administration. *Atmos Environ* 43:4472–4480
  146. Gertler A, Kuhns H, Abu-Allaban M, Damm C, Gillies J, Etyemezian V, Clayton R, Proffitt D (2006) A case study of the impact of winter road sand/salt and street sweeping on road dust re-entrainment. *Atmos Environ* 40(31):5976–5985
  147. Chang Y-M, Chou C-M, Su K-T, Tseng C-H (2005) Effectiveness of street sweeping and washing for controlling ambient TSP. *Atmos Environ* 39:1891–1902
  148. Chou C-M, Chang Y-M, Lin W-Y, Tseng C-H, Chen L (2007) Evaluation of street sweeping and washing to reduce ambient PM<sub>10</sub>. *Int J Environ Pollut* 31(34):431–448
  149. Kuhns H, Etyemezian V, Landwehr D, MacDougall C, Pitchford M, Green M (2001) Testing re-entrained aerosol kinetic emissions from roads (TRAKER): a new approach to infer silt loadings on roadways. *Atmos Environ* 35:2815–2825
  150. John AC, Hugo A, Kaminski H, Brandt A, Kappert W, Falkenberg E, Kuhlbusch TAJ (2007) Effectiveness of street cleaning for reducing ambient PM<sub>10</sub> concentrations. In: European aerosol conference 2007, Salzburg, Abstract T19A047
  151. Amato F, Querol X, Alastuey A, Pandolfi M, Moreno T, Gracia J, Rodriguez P (2009) Evaluating urban PM<sub>10</sub> pollution benefit induced by street cleaning activities. *Atmos Environ* 43:4472–4480
  152. Berthelsen B-O (2003) The use of magnesium chloride as dust reducer at E6 through Trondheim (only in Norwegian). Report No. TM2003/2, Trondheim Municipality, Environment Division, Trondheim, ISBN:82-7727-087-9
  153. Johansson C, Burman B, Forsberg B (2009) The effects of congestions tax on air quality and health. *Atmos Environ* 43:4843–4854
  154. Bowker GE, Baldauf R, Isakov V, Khlystov E, Petersen W (2007) The effects of roadside structures on the transport and dispersion of ultrafine particles from highways. *Atmos Environ* 41:8128–8139
  155. Baldauf R, Thoma E, Khlystov A, Isakov V, Bowker G, Long SR (2008) Impacts of noise barriers on near-road air quality. *Atmos Environ* 42:7502–7507
  156. Morikawa H, Takahashi M, Hakata M, Sakamoto A (2003) Screening and genetic manipulation of plants for decontamination of pollutants from the environment. *Biotechnol Adv* 22:9–15
  157. Beckett KP, Freer-Smith P, Taylor G (2000) Effective tree species for local air quality management. *Journal of Arboriculture* 26(1):12–19
  158. Yang J, Yu Q, Gong P (2008) Quantifying air pollution removal by green roofs in Chicago. *Atmos Environ* 42:7266–7273
  159. Cavanagh JE (2006) Potential of vegetation to mitigate road-generated air pollution. Part I – review of background information. Contractor Report LC0506/00xx. Landcare Research, Lincoln
  160. Litschke T, Kuttler W (2008) On the reduction of urban particle concentration by vegetation – a review. *Meteorol Z* 17(3):229–240
  161. Laufs S, Burgeth G, Duttlinger W, Kurtenbach R, Maban M, Thomas C, Wiesen P, Kleffmann J (2010) Conversion of nitrogen oxides on commercial photocatalytic dispersion paints. *Atmos Environ* 44:2341–2349
  162. Child & Associates (2004) M5 East Freeway: A review of emission treatment technologies, systems and applications. Review undertaken by Child & Associates for the Roads and Traffic Authority of NSW. [http://www.rta.nsw.gov.au/constructionmaintenance/downloads/2004\\_10\\_childrep-filtration.pdf](http://www.rta.nsw.gov.au/constructionmaintenance/downloads/2004_10_childrep-filtration.pdf)
  163. Willoughby P, Stricker J, Humphrey G (2004) Electrostatic precipitators and ventilation in road tunnels in Japan. Report of a visit by a delegation from the NSW Roads and Traffic Authority to Japan from 30 September – 10 October 2003. New South Wales Roads and Traffic Authority
  164. TNO (2009) Conceptrapport IPL. Praktijkonderzoek “Elektrostatisch concept in Thomassentunnel”. Report of the Dutch Air Quality Innovation Programme (IPL)
  165. Davis SC, Diegel SW, Boundy RG (2010) Transportation energy data book: edition 29. Report ORNL-6985. Oak Ridge National Laboratory, Oak Ridge
  166. Eyring V, Stevenson DS, Lauer A, Dentener FJ, Butler T, Collins WJ, Ellingsen K, Gauss M (2007) Multi-model simulations of the impact of international shipping on atmospheric chemistry and climate in 2000 and 2030. *Atmos Chem Phys* 7:757–780
  167. Jalkanen JP, Brink A, Kalli J, Pettersson H, Kukkonen J, Stipa T (2009) A modelling system for the exhaust emissions of marine traffic and its application in the Baltic Sea area. *Atmos Chem Phys* 9:9209–9223
  168. Smit R, Ntziachristos L, Boulter P (2010) Validation of road vehicle and traffic emission models – a review and meta-analysis. *Atmos Environ* 44:2943–2953

## Books and Reviews

- Boulter P, McCrae I (eds) (2009) ARTEMIS: assessment and reliability of transport emission models and inventory systems. Final Report. TRL Report PPR350. TRL Limited, Wokingham
- Finlayson-Pitts BJ, Pitts JN (2000) Chemistry of the upper and lower atmosphere. Academic, New York. ISBN 0-12-257060-X
- Sokhi RS (ed) (2008) World atlas of atmospheric pollution. Anthem Press, London. ISBN 9781843312895
- USEPA (2009) Integrated science assessment for particulate matter. EPA/600/R-08/139F. United States Environmental Protection Agency. <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=216546>
- Wayne RP (2000) Chemistry of atmospheres, 3rd edn. Oxford University Press, Oxford, UK. ISBN 0-19-850375-X

## Airborne and Space-borne Remote Sensing of Cryosphere

KENNETH C. JEZEK

Byrd Polar Research Center, School of Earth Sciences,  
The Ohio State University, Columbus, OH, USA

### Article Outline

Glossary  
Definition of the Subject: The Cryosphere  
Introduction  
Early History of Cryospheric Remote Sensing  
Scientific Advances from Airborne and Space-borne Remote Sensing  
Cooperative Efforts to Observe, Monitor, and Understand the Cryosphere  
Future Directions  
Bibliography

### Glossary

**Cryosphere** Those components of the Earth system that contain water in its frozen form.

**Radar** Radio detection and ranging systems.

**Lidar** Light detection and ranging systems.

**Radiometers** Radio frequency receivers designed to detect emitted radiation from a surface and in accordance with Planck's law.

**Synthetic aperture radar** Radar system which increases along track resolution by using the motion of the platform to synthesize a large antenna.

**Permafrost** Persistently frozen ground.

**Ice sheet** Continental-scale, freshwater ice cover that deforms under its own weight.

**Sea ice** Saline ice formed when ocean water freezes.

**Glaciers** Long, channelized, slabs of freshwater ice thick enough to deform under their own weight.

**Seasonal snow** The annual snow that blankets land cover in the winter and melts by summer.

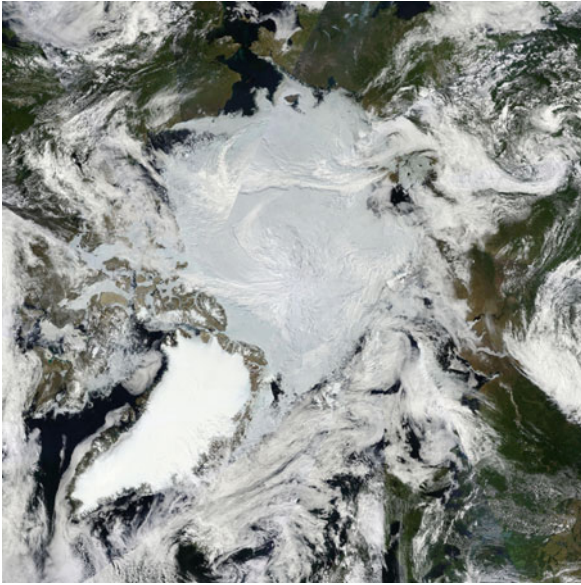
### Definition of the Subject: The Cryosphere

The Cryosphere broadly constitutes all the components of the Earth system which contain water in a frozen state [1]. As such, glaciers, ice sheets, snow cover, lake and river ice, and permafrost make up the terrestrial elements of the Cryosphere. Sea ice in all of its forms, frozen sea bed and icebergs constitute the oceanic elements of the Cryosphere while ice particles in the upper atmosphere and icy precipitation near the surface are the representative members of the Cryosphere in atmospheric systems. This overarching definition of Earth's cryosphere immediately implies that substantial portions of Earth's land and ocean surfaces are directly subject in some fashion to cryospheric processes. Through globally interacting processes such as the inevitable transfer of heat from the warm equatorial oceans to the cold polar latitudes, it seems reasonable to argue that all regions of Earth are influenced by cryospheric processes and their integration into the modern climate of the planet. Observations of the cryosphere necessary to predict future variability in Earth's ice cover and its interaction with other Earth systems must be made on commensurate spatial and temporal scales. Consequently, airborne and space-borne remote sensing technologies with global reach play a key role in acquiring data necessary to understand the important physical processes and earth system interactions that govern the evolution of the Cryosphere (Fig. 1).

### Introduction

The broad spatial and seasonally changing distribution of ice plays an important role in earth systems and human activities. At high latitudes, ice covered land and ocean surfaces are highly reflective thus redirecting incoming solar radiation back into space in the summer months. Indeed, reductions in the spatial area of snow and ice cover are believed to be an important





**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 1**

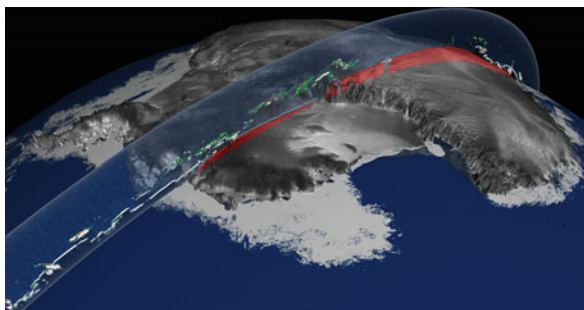
NASA MODIS composite for July 7, 2010. Starting in the lower left quadrant and going clockwise, the Greenland sheet is almost cloud free as are the adjacent smaller ice caps on the Canadian Arctic Islands. Much of northern Canada and Alaska are cloud covered till the Bering Strait which is also almost sea ice free. Patches of sea ice cling to the Asian coast and there is a smattering of snow cover on the Putorana Mountains in western Siberia. The Gulf of Ob, Novaya Zemlya, the Franz Josef Land, and Spitsbergen Islands are barely visible through the cloud. The central Arctic basin is covered by sea ice. The image illustrates the scale of the cryosphere as well as some of the complications faced when trying to study it from space (Image prepared by NASA's MODIS Rapid Response team)

feedback mechanism that enhances warming at high latitudes [2, 3]. Essentially, reduced snow and ice cover exposes darker land and ocean surfaces that retain rather than reflect solar energy. This results in increased warming and hence a further decrease in the area of snow and ice covered surfaces. Sea ice is an important habitat for birds such as penguins and mammals such as seals and polar bears which thrive in this icy environment [4]. But the underside of sea ice is also an important refuge for some of the smallest creatures including the shrimp-like krill which graze upon algae that grows just beneath and within the ice canopy

[5, 6]. Terrestrial permafrost and frozen sediments beneath the oceans support an important reservoir of organic carbon and gas hydrates [7]. As permafrost melts, methane can be released contributing to the increasing concentration of greenhouse gases in the atmosphere. Glaciers and ice sheets are vast reservoirs of Earth's freshwater. As the ice sheets thin, water flows from the ice sheets into the oceans raising global sea level [8]. In terms of our daily activities, seasonal snow and glaciers are important sources of spring runoff for irrigation and power generation, while ice jams on rivers constitute important obstacles to winter-time navigation and can cause low-land flooding [9]. Thawing permafrost causes the land near surface to become unstable which can result in catastrophic structural failures in buildings.

The global span of cryospheric processes and the strong daily to seasonal swing in the extent of snow and ice makes studying and monitoring the cryosphere an especially challenging scientific objective [10]. Moreover the variety of forms in which ice can be manifest in Earth systems means that no single observing system is capable of making adequate observations. Rather an ensemble of techniques is required to fully appreciate and eventually understand the complexities of the cryosphere and its interaction with other earth systems. Locally, observing tools may include direct, field measurements of snow pack thickness or physical temperature. Much different sets of tools are needed to characterize the cryosphere on a global and annual scale where the inhospitable climate and the physical remoteness of many sectors of the cryosphere represent obstacles to scientific investigation. Here, aircraft and spacecraft mounted instruments are required to infer geophysical properties from a remote distance (Fig. 2).

Compounded with these geographic challenges, observations of the icy surface often have to overcome cloud cover that frequently obscures the high latitudes, as well as continuing measurements during the long polar night. Consequently a range of instruments generally rely on distinguishing properties evident in the broad-spectral electrical characteristic of icy terrain. For example, even the cloudy atmosphere is largely transparent to microwave radiation. That fact combined with the very different microwave emission of open ocean and sea ice enables space-borne microwave observations of annual sea ice extent and concentration



### Airborne and Space-borne Remote Sensing of Cryosphere. Figure 2

Illustration of NASA's ICESat passing over the Antarctic. ICESat collected repeat elevation data over ice sheets and sea ice to study changing ice thicknesses. ICESat also sounded the atmosphere to study aerosols. NASA/Goddard Space Flight Center Scientific Visualization Studio, RADARSAT mosaic of Antarctica (Canadian Space Agency)

day and night and in all weather. Similarly the electrical contrast between rock and ice enables airborne radio frequency radar measurements of the thickness of polar ice sheets and glaciers. However some of the most recent advances in cryospheric science have been made by relying on one of the most basic properties of the icy cover, namely its mass. Space-borne measurements of the changing gravitational attraction of variably sized snow and ice bodies enables direct estimates of changing polar ice mass and the redistribution of melting ice into the liquid oceans. Indeed it can be argued that engineering and technological advances in airborne and remote sensing have had some of their greatest scientific impacts in the understanding the evolution of Earth's ice cover.

This article provides an overview of remote sensing of the cryosphere from both aircraft and spacecraft. A brief historical review of remote sensing of snow and ice is followed by a discussion of the physics of remote sensing of the cryosphere and how developments in remote sensing have led to a series of important scientific advances. These include the realization of the diminishing extent of Arctic sea ice and the thinning of glaciers, ice caps, and ice sheets worldwide. Both of these observations form critical, direct evidence of changing world climate. The article concludes with a discussion of developing international collaborations

that are aimed at pooling technologically sophisticated and operationally expensive international assets so as to obtain an integrated system of continuing remote sensing observations necessary to predict future changes in Earth's ice and the consequent impacts on human activities.

### Early History of Cryospheric Remote Sensing

The science and operational communities have always been quick to adopt new instruments and platforms for use in observing snow and ice in all of its forms. This section reviews early developments in cryospheric remote sensing – which were in part driven by science and in part by the basic desire to explore the most remote regions of Earth.

### Airborne Photo-Reconnaissance

Aerial photography of ice covered terrain began during early twentieth century expeditions to the high altitudes and was used primarily to document the progress of the expedition. Survey quality aerial mapping was adopted using techniques primarily developed during World War I [11]. Mittelholzer and others [12] writing about the 1923 Junkers Expedition to Spitzbergen offer a very complete overview of the geographic and cartographic objectives for aerial photography in the North including a brief discussion of glacier formation as revealed by the aerial photographs. They also give an interesting technical discussion of the challenges faced when doing aerial photographic reconnaissance over highly reflective snow-covered terrain.

Wilkins documented ice cover in the Antarctic Peninsula during the first successful flight in Antarctica by using a hand-held, folding Kodak 3A camera [13, 14]. Richard E. Byrd devoted time and resources to aerial photography for quantitative surveying purposes during his first Antarctic Expedition of 1928–1930. In his book, "Little America," Byrd [15] writes that photographs from Ashley McKinley's laboratory provided "perhaps, the most important geographical information from the expedition." McKinley was third in command of the expedition and the aerial surveyor. McKinley operated his Fairchild K-3 mapping camera during Byrd's 1929 historic flight to the South Pole.

These early airborne photographic records were acquired with great skill and at considerable risk. The

quality of the photographs is often exceptional and the photos themselves represent an oft under-utilized resource for directly gauging century-scale changes in Earth's ice cover.

### Satellite Photography

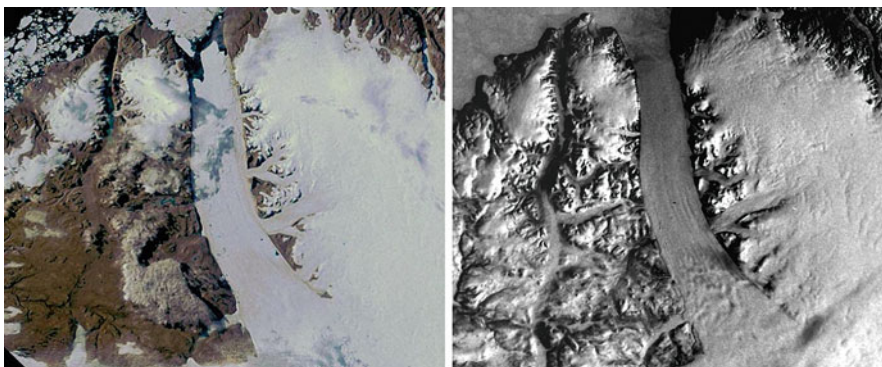
A mere 32 years after Byrd's aerial photographic surveying in Antarctica, space-borne cameras began capturing unique pictures of Earth. CORONA, ARGON, and LANYARD were the first three operational imaging satellite reconnaissance systems and they acquired data during the early 1960s for both detailed reconnaissance purposes and for regional mapping [16–18]. Early reconnaissance satellite photographs provide a unique view of our world as it appeared at the beginning of the space age. Researchers in the environmental science community are the most recent beneficiaries of these data after they were declassified and made publicly available in 1995 through the efforts of Vice President Al Gore along with several government agencies working together with civilian scientists as part of the MEDEA program [19]. Polar researchers in particular inherited a wealth of detailed photography covering both of the great polar ice sheets. After processing with sophisticated digitizing instruments and subsequent analysis with modern photogrammetric and image processing techniques, investigators have shown that these data can be used to characterize local fluctuations in glacier termini

[20, 21], investigate large-scale flow features on ice sheets [22, 23] and to measure long-term average velocity by feature retracking techniques [24] (Fig. 3).

### Development of Depth Sounding Radar

Early suggestions that glaciers were penetrated by radio signals are attributed to observations made at Little America during Byrd's 2nd Antarctic Expedition [25]. This observation, along with reports that pulsed radar altimeters were yielding faulty readings over glaciers, led to the first radar experiments to measure ice thickness in 1955. In 1960, Waite and Schmidt [25] made measurements over Greenland from aircraft at 110, 220, 440, and 4,300 MHz. These results initiated a revolution in glaciology because the ice thickness and internal structure of glaciers and ice sheets could be rapidly sounded from aircraft [26].

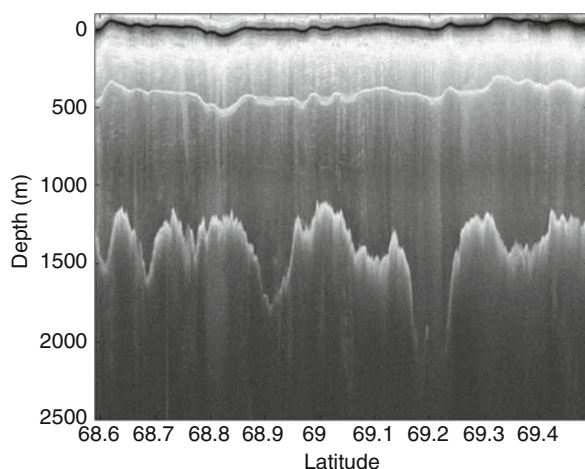
Although the sophistication of ice sounding radar systems has increased tremendously, the basic principle of the technique remains the same and continues to be a fundamental tool used by researchers (Fig. 4). Essentially, the one-way travel time of a radar pulse transmitted through the ice or snow is multiplied by the appropriate wave speed and the thickness so determined. More complex, two dimensional maps of the glacier bed topography can be assembled from multiple profiles that are combined using travel time migration techniques [27]. Several snow and ice sounding radars are part of the primary instrument suite presently



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 3**

European Space Agency MERIS image of Peterman Glacier, Greenland (*left*). The multispectral image was acquired in July, 2008 and before a large part of the floating ice tongue broke free. Argon panchromatic photograph acquired during the spring of 1962. The clarity of this early view from space is exceptional and provides a valuable gauge for assessing changes in glacier ice from the 1960s to the present





**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 4**

North-south airborne profile of ice thickness in west central Greenland. The *upper black line* is the radar reflection from the snow surface. The line at about 500 m depth is a multiple echo of the surface from the aircraft itself. The undulating line at an average depth of about 1,400 m is the reflection from the glacier bottom. The profile crosses the upstream segment of Jakobshavn Glacier at 69.2° N 48.1°W. There the 2,000 m thick ice has incised a deep channel into the bedrock. The data were acquired by the University of Kansas in 2008

carried aboard aircraft supporting NASA's IceBridge program [28].

Depth sounding radar remains one of the few techniques available to researchers interested in probing the volume of the terrestrial ice cover and the properties of the underlying bed. Airborne radars have been successfully used to study glacier, ice sheets, and to a more limited degree sea ice and permafrost. Seismic techniques yield important complementary information but can be carried out only *in situ*. Airborne gravity measurements provide important regional information but are generally less accurate than the radar technique for measurements on glaciers.

### Scientific Advances from Airborne and Space-borne Remote Sensing

In most instances, airborne and space-borne platforms carry similar kinds of remote sensing instruments. Often times, space-borne instruments are first proto-

typed as part of initial airborne campaigns. However, advantages of instrument installations on multiple kinds of platforms go far beyond vetting the effectiveness of an instrument. For example, space-borne systems can provide global scale observations, with some imaging instruments yielding pole-to-pole observations on a daily basis. Manned and increasingly unmanned aircraft, with their ability to maneuver and loiter over an area, can provide much denser spatial and temporal coverage when profiling instruments, such as altimeters, are of interest. This section begins with a review of remote sensing physics and then summarizes key applications of remote sensing to several cold regions themes.

### Basics of Remote Sensing of the Cryosphere

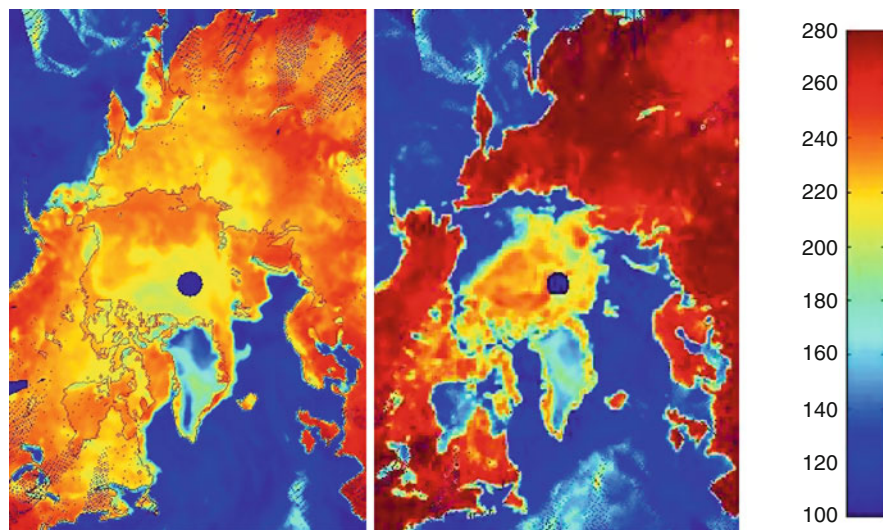
Airborne and space-borne sensors measure local changes in electromagnetic, gravitational, and magnetic force fields. Disturbances can arise from the passive, thermodynamically driven electromagnetic radiation emitted by the earth's distant surface, or the reflectance of incoming solar radiation from the Earth's surface, both of which change with the terrain type. Instruments and associated satellites of this type include the Electrically Scanning Microwave Radiometer (ESMR), Scanning Multichannel Microwave Radiometer (SMMR), Special Sensor Microwave Radiometer (SSM/I), Advanced Very High Resolution Radiometer (AVHRR), Landsat, Moderate Resolution Imaging Spectrometer (MODIS), Medium Resolution Imaging Spectrometer (MERIS), Satellite Pour l'Observation de la Terre (SPOT). Disturbances can also arise from the active behavior of the sensor itself. Radars and lidars include the Airborne Topographic Mapper (ATM), Laser Vegetation Imaging Sensor (LVIS), synthetic aperture radar (SAR) and radar altimeters on the European Remote Sensing Satellite (ERS-1/2), RADARSAT 1/2, synthetic aperture radar and radar altimeter on Envisat, Phased Array L-band SAR, TerraSAR-X, GEOSAT, the Ice and Climate Experiment Satellite (ICESat) lidar, the Cryosat radar altimeter. These instruments illuminate the surface with electromagnetic signals, which upon reflection can be detected back at the sensor after an elapsed time. Local changes in the gravity field at airborne and space-borne elevations are indications of changes in the distribution of mass within the Earth. Such instruments include airborne

gravimeters on NASA's IceBridge, NASA's Gravity Recovery and Climate Experiment (GRACE) satellite, and ESA's Gravity field and Ocean Circulation Explorer (GOCE) satellite.

The task of the remote sensing scientist is to take measurements of these basic force fields and infer from them geophysical properties about the Earth [29]. For the cryosphere, examples include using changes in the gravitational field to estimate changes in the mass of glaciers and ice sheets as is done with the GRACE satellite. Other examples include using radars and lidars to measure the time of flight of signals reflected off the ice surface and use that information to measure elevation and elevation changes on sea ice and ice sheets. Finally, scattered solar radiation can be used to acquire hyper-spectral images for surface characterization and for traditional mapping of surface features.

Accomplishing this task requires technical information about the behavior of the instrument, knowledge about the physical relationships between the measured field and the surface under study, and the development of algorithms that correctly invert the measured signal into some desired property [30, 31]. In some cases this

task can be relatively easy. Because glacier ice is nearly homogeneous and because it is almost transparent at radar frequencies from about 1 to 500 MHz [32], the propagation time of a radar echo through the ice sheet and back can be converted to ice thickness by simply knowing the average ice dielectric constant, which in turn is readily convertible into a propagation velocity. In other cases, the task is more challenging because of heterogeneities in the material. The physical and electrical properties of sea ice change substantially as the ice ages [33]. This leads to changes in the thermally driven, microwave emission from the surface. The amount of energy received, measured in terms of a brightness temperature, is related to the product of the physical temperature and the emissivity of the surface. For open water, the microwave brightness temperature is cool because little energy escapes from the ocean surface. For sea ice, the brightness temperature is warm because more energy is transmitted across the electrically less reflective snow surface at these frequencies. Consequently, brightness temperature can be used to measure sea ice concentration and extent with high accuracy (several percent). Figure 5 shows that just on the basis



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 5**

19 GHz, horizontally polarized, special sensor microwave imager (SSM/I) brightness temperature data for the Arctic on February 15, 2004 (*left*) and August 25, 2004 (*right*). The brightness temperature scale (*far right*) is in degrees Kelvin. Differences between the emissivity of ocean, sea ice, land and ice sheets enable the relatively easy identification of the seasonal retreat in sea ice cover. Brightness temperature variations across the ice pack are caused by differences in ice type and age as well as in ice concentration



of brightness temperature maps alone it is easy to distinguish the annual cycle of sea ice growth and decay across the arctic.

### Sea Ice Extent, Concentration, Motion, and Thickness

Sea ice modulates polar climate by restricting the flow of heat from the relatively warm polar ocean into the relatively cold polar atmosphere. Because freezing ice preferentially rejects impurities from the crystalline lattice, the growth of sea ice modulates ocean circulation by releasing dense, cold brine that sinks beneath the marginal ice zones and initiates oceanic convection. Sea ice represents a natural barrier to surface navigation and so changes in sea ice cover have important consequences for future development of the Arctic. While sea ice is usually distinguishable from open ocean at optical wavelengths, cloud cover and the long polar night dictate the use of all-weather, day/night passive microwave radiometers for monitoring sea ice extent and concentration. As noted above, these radiometers measure the radiant energy emitted from the surface that, on one hand, can help identify ice type and age but, on the other hand, leads to complexities in designing more sophisticated instruments capable of sorting out whether the observed changes are due to different ice types populating a scene or because of change in the fractions of sea ice and open water in the scene. Moreover, the dimensions of the area sampled on the surface are often quite large (Hollinger and others [34] quote a field of view of  $37 \times 28$  km for the SSMI 37 GHz vertically polarized channel). Consequently, approaches for estimating the concentration within a single pixel are required. A simple algorithm relies on the fact that the total emitted energy or equivalently the brightness temperature must equal the brightness temperature of the components in the scene times their respective concentrations within the pixel [35]. The component brightness temperatures are selected by tie points (such as a tie point for the brightness temperature of open water or first-year sea ice) which allows for a solution. More sophisticated algorithms attempt to better characterize regions where ice concentrations are low and where the data may be contaminated by weather effects. The results of these analyses are detailed measurements of ice extent which

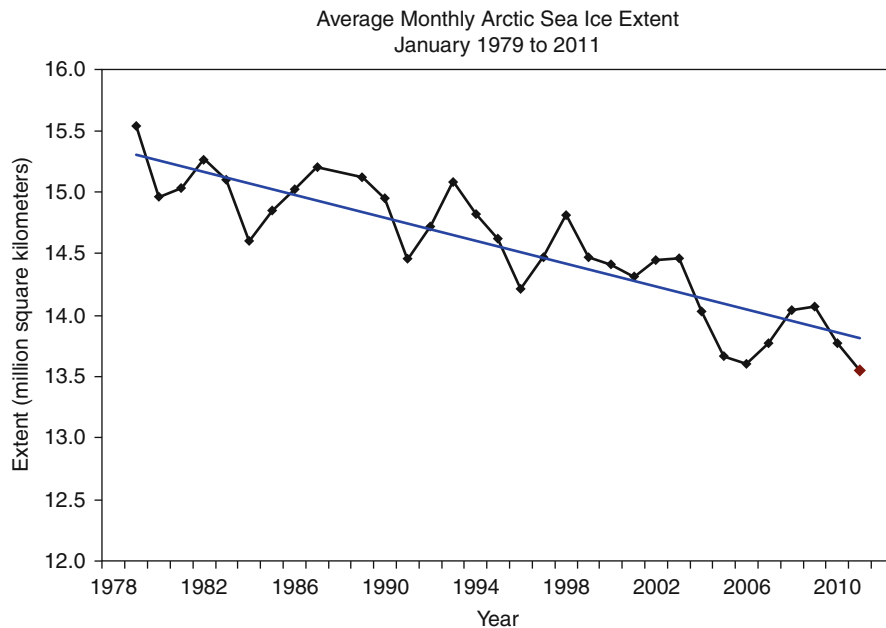
document the dramatic decrease in Arctic sea ice cover (Fig. 6). Graphs such as these represent some of the most compelling and straightforward evidence for changing climate at high latitudes.

Images such as those in Fig. 5 can also be used to document the coarse motion of the sea ice by tracking common features in repeat images. Better results are obtained with increasing resolution ranging from the several kilometer resolution obtainable with AVHRR to the very fine resolutions (tens of meters or less) achievable with SAR. Combined with airborne (Airborne Topographic Mapper flown as part of IceBridge) and space-borne (ERS-1/2, ICESat, Cryosat-2) altimeter estimates of ice thickness, circulation driven changes in ice thickness about Antarctica [36] and the total ice flux across the Arctic and out into the marginal seas can be computed. Based on ICESat estimates of sea ice freeboard, Kwok and others [37] estimate a 0.6 m thinning in Arctic multiyear ice for the period from 2003 to 2008.

### Regional Image Mapping of Glaciers and Ice Sheets

Airborne and space-borne image mapping of glaciers and ice sheets focuses on basic physical characteristics such as glacier termini, glacier snow lines, crevasse patterns, snow facies boundaries [38]. These properties have been successfully monitored with optical and microwave imaging instruments. As described in section “Early History of Cryospheric Remote Sensing,” high latitude mapping began as early as 1962 with the launch of the Argon satellite. Shortly thereafter, the 1970s Landsat 1, 2, and 3 Multi-Spectral Scanner (MSS) images constitute an important glaciological resource [39, 40] and have been compiled into a series of beautiful folios edited by R. Williams Jr. and J. Ferrigno of the U.S. Geological Survey [see 41].

Compiling separate images into seamless, high-resolution digital maps of continental-scale areas is complicated by the data volume and the challenges in accurately estimating satellite orbits and instrument viewing geometries along orbit segments that might span from coast to coast. Initial successes in large-scale mapping were achieved through use of the moderate spatial resolution (1–2.5 km) and wide swath (2,400 km) Advanced Very High Resolution Radiometer (AVHRR) images [42] which helped reveal details



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 6**

Winter-time Arctic sea ice extent computed using passive microwave data. The decreasing trend illustrates the reduction in Arctic sea ice extent. There is a much more dramatic decline in ice extent during the summer months (Graph courtesy of the National Snow and Ice Data Center)

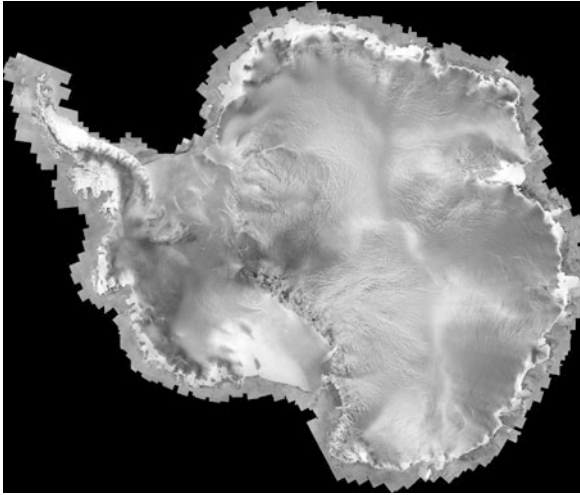
about ice stream flow in West Antarctica [43]. After the original AVHRR mosaic of Antarctica, the United State Geological Survey (USGS) made subsequent improvements to the mosaic by eliminating more cloud, separating the thermal band information to illustrate surface features more clearly, and correcting the coast-line of the mosaic to include grounded ice while excluding thin, floating fast ice [44]. The European Earth Resources Satellite – 1 carried on board a synthetic aperture radar which allowed for large-scale regional mapping. Fahnestock and others [45] compiled a mosaic of Greenland which revealed the existence of a long ice stream in north east Greenland. In 1997, RADARSAT-1 synthetic aperture radar (SAR) data were successfully acquired over the entirety of Antarctica (Fig. 7). The coverage is complete and was used to create the first, high-resolution (25 m) radar image mosaic of Antarctica [46–48]. The image was used to map the ice sheet margin [49] and to investigate patterns of ice flow across Antarctica resulting in the discovery of large ice streams that drain from Coates Land into the Filchner Ice Shelf. Other large-scale mapping has been completed with MODIS for both the

Arctic and the Antarctic [50]. Most recently, Landsat imagery of Antarctica has been compiled into a single, easily accessible map-quality data set [51] and SPOT stereo imagery has been used to derive digital elevation models of ice sheets, ice caps, and glaciers [52].

### InSAR Measurements of Glaciers and Ice Sheets Surface Velocity

Glaciers and ice sheets move under the load of their own weight. They spread and thin in a fashion dictated by their thickness, the material properties of ice, and the environmental conditions operative on the glacier surface, sides and bottom. The rate and direction of motion reveals important information about the forces acting on the glacier, provides knowledge about the rate at which ice is pouring into the coastal seas, and enables scientists to predict how the ice sheet might respond to changing global climate.

Since the International Geophysical Year of 1957–1958 and before, scientists have placed markers on the ice sheet and then, using a variety of navigation techniques from solar observations to GPS, have remeasured



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 7**

Canadian RADARSAT-1 Antarctic Mapping Project synthetic aperture radar mosaic. The radar is sensitive to changes in the physical structure of the near surface snow. Bright coastal returns are scattered from subsurface ice lenses formed during fall freeze-up. Darker interior tones occur where snow accumulation is high. Other features such as ice streams and ice divides are visible (Map courtesy of K. Jezek. RADARSAT-1 data courtesy of the Canadian Space Agency)

their positions to calculate motion. More recently, scientists have used high-resolution satellite images to track the position of crevasses carried along with the glacier to compute surface motion. These approaches are time consuming and result in patchy estimates of the surface velocity field. During the early 1990s, researchers at the Jet Propulsion Laboratory showed that synthetic aperture radar (SAR) offered a revolutionary new technique for estimating the surface motion of glaciers [53]. Here, the SAR is operated as an interferometer. That is, the distance from the SAR to a point on the surface is computed by measuring the relative number of radar-wave cycles needed to span the distance between the radar and the surface. Later, another measurement is made from a slightly different position and the numbers of cycles is computed again. The difference in the number of cycles combined with control points is used to estimate relative displacement to about one quarter of a radar-wave cycle (just a few centimeters for RADARSAT-1).

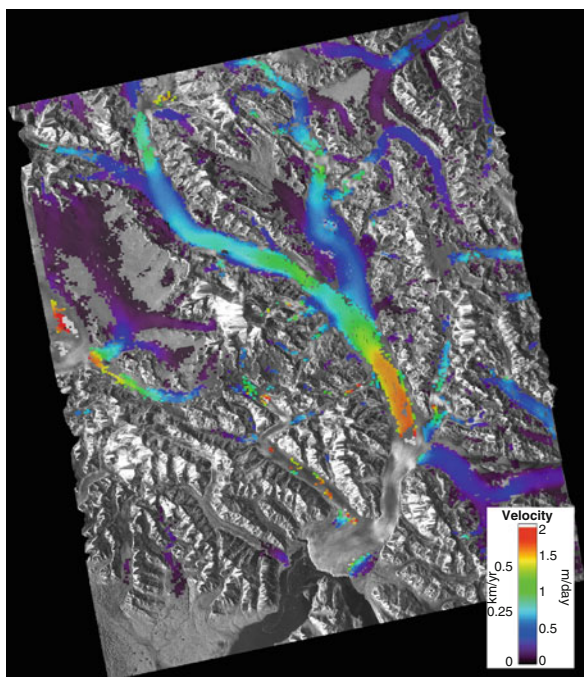
Given an estimate of surface topography, this enables measurement of even the slowest moving portions of glaciers [54, 55]. A related approach relies on the fact that images formed from coherent radar signals scattered off a rough surface will have a small-scale-speckly appearance. The speckle pattern is random over scene but it is stable from scene to scene over a short time. Tracking the speckle pattern over time enables another measurement of surface velocity and while less accurate than the interferometric phase approach, speckle retracking has the advantage of yielding estimates of two components of the velocity vector [56, 57]. Typically a mixture of phase interferometry and speckle retracking are used in glacier motion studies.

Surface topography can also be estimated using InSAR techniques in cases where the surface velocities are small, the period between repeat observations is short [58], when multiple interferometric pairs are available [59], or when multiple antennas enable acquisition of interferometric data in a single pass as was done with NASA's Shuttle Radar Topographic Mission [60]. Although the surface elevation accuracy of InSAR topography is typically on the order of a few meters, the results can be used to estimate local slopes and also in comparison with earlier data provide a rough estimate of mass change.

Surface elevation and velocity data are key to studies of glacier dynamics [61]. InSAR data have been used to investigate the relative importance of different resistive stresses acting on ice streams and to predict the future behavior of ice streams and glaciers currently in retreat [62, 63]. Measurements on mostly retreating glaciers in the Himalayas [64], Patagonia [65, 66], European Alps [67], and Alaska (Fig. 8) document changing surface elevation and internal dynamics, and, together with estimates of ice thickness and surface accumulation rate, can provide a direct estimate of glacier mass loss [68].

### Glaciers and Ice Sheets Mass Loss

Glaciers and ice sheets are reservoirs of freshwater with over 90% of Earth's freshwater bound in the Antarctic Ice Sheet [69], which when depleted have local, regional and global impacts. Indirect approaches for identifying whether ice sheets are losing mass include using proxy indicators such as surface melt



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 8**

Surface velocities on Hubbard Glacier Alaska measured using Japanese PALSAR L-band interferometric pairs. Hubbard Glacier is a tidewater glacier and ice from the advancing snout calves directly into Disenchantment Bay (*lower center*). Hubbard Glacier is one of the few in Alaska that is currently thickening and advancing (Image courtesy of E. W. Burgess and R. R. Forster, University of Utah)

area and duration – both measureable using passive microwave techniques [70, 71]. More directly, there are three primary remote sensing techniques currently used to assess the changing volume (or mass) of ice contained in glaciers [72]. The first involves an estimate of the difference between the annual net accumulation of mass on the surface of the glacier and the flux of ice lost from the terminus. The flux from the terminus is calculated using the measured ice thickness from airborne radar and the surface velocity, which is currently best estimated with InSAR [73]. The second approach is to measure surface elevation change. This has been accomplished with space-borne radar altimeters [74, 75] and airborne and space-borne lasers [76–78]. Finally, the changing glacial mass can be estimated directly by measuring changes in the gravity field as has been done with GRACE [79–81]. **Figure 9**

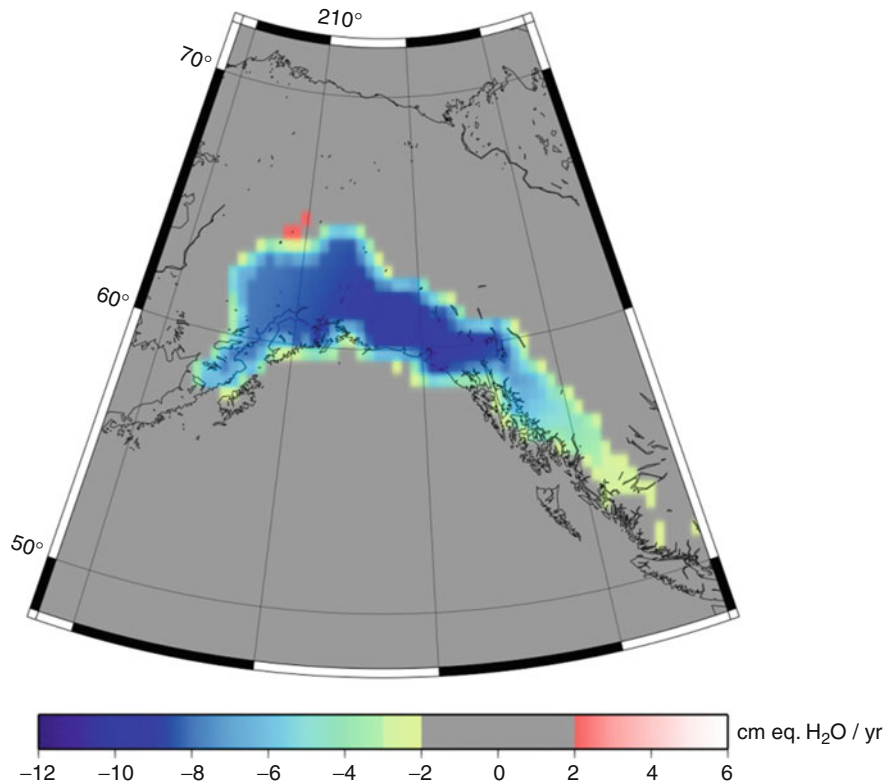
illustrates the estimated mass loss rate from Alaskan Glaciers as illustrated using GRACE data (Updated from [82]).

Recent analyses suggest these different techniques are yielding similar mass reductions for the Greenland Ice sheet [8, 83]. Thinning is primarily in the coastal regions where the rate of thinning has been increasing since the early 1990s and the current rates of mass loss estimated between 150 and 250 Gt/year. West Antarctica data indicate mass loss similar to Greenland, whereas East Antarctica retains a slightly positive (50 Gt/year) mass balance. Using surface mass balance estimates and GRACE gravity data, Rignot and others [84] report a combined ice sheet thinning rate of  $475 \pm 158$  Gt/year. Cazenave and Llovel [8] conclude that for the period between 1993 and 2007 about 55% of total sea level rise can be attributed to melting of glaciers and ice sheets.

### The Seasonal Snow Pack

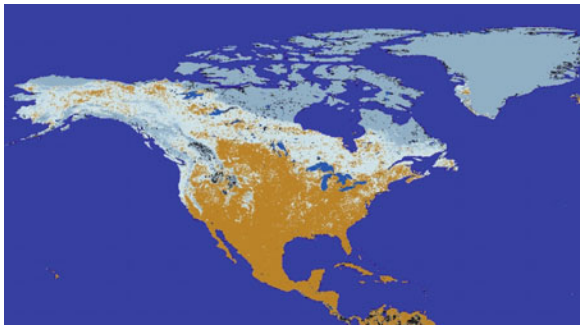
Seasonal snow cover plays an important role in regional hydrology and water resource management. Rapid melting of the seasonal snow pack across the northern Great Plains in April 1997 resulted in catastrophic flooding of the Red River. From a climate perspective, the bright snow surface also serves as an effective mirror for returning incoming solar radiation back into space thus modulating the planetary heat budget.

The spectral reflectivity differences between snow, cloud, and other land cover types enables the seasonal snow cover to be routinely mapped globally using visible and infrared sensors such as NOAA's AVHRR, NASA's MODIS, and ESA's MERIS instruments. Snow may be mapped based on visual inspection of multi-spectral imagery. Automatic snow detection is accomplished by computing the normalized difference between, for example, the MODIS visible band (0.545–0.565  $\mu\text{m}$ ) and the near infrared band (1.628–1.652  $\mu\text{m}$ ). Snow is detected when the normalized difference exceeds a threshold value of 0.4 and when other criteria on land and cloud cover are met [85] (Fig. 10). Based on analysis of the NOAA 35 year data record, Dery and Brown [86] conclude that springtime snow extent across the northern hemisphere declined by some  $1.28 \times 10^6 \text{ km}^2$  over a 35 year period.



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 9**

Mass loss from Alaskan Glaciers in cm of water equivalent per year from GRACE (Image courtesy of S. Luthcke, NASA GSFC)



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 10**

Snow cover extent from NASA's MODIS. Increasing shades of *gray* indicate greater percentages of snow cover within each pixel. *Brown* denotes snow-free surface

Snow thickness and hence indirectly the mass of snow are key variables for estimating the volume of water available in a reservoir and potentially releasable as runoff. The most successful techniques to date have

relied on passive-microwave-based algorithms. One approach for estimating snow thickness is to difference 19 and 37 GHz brightness temperature data that along with a proportionality constant yields an estimate of the snow thickness. The algorithm is based on the fact that 19 GHz radiation tends to minimize variations in ground temperature because it is less affected by the snow pack. The 37 GHz radiation is strongly scattered by the snow grains and brightness temperature at this frequency decreases rapidly with snow thickness/snow water equivalent. Factors which confuse this algorithm include topography and changes in ground cover [87].

Information about the seasonal onset of snowmelt can be obtained from microwave data. A few percent increase in the amount of free water in the snow pack causes the snow emissivity to approach unity resulting in a dramatic increase in passive microwave brightness temperature. This fact has been successfully used to track the annual melt extent on the ice sheets and also to track the springtime melt progression across the



arctic. Higher resolution estimates of melt extent can be obtained with scatterometer and SAR data. These data generally show an earlier spring time date for the beginning of melt onset and a later date for the fall freeze [88].

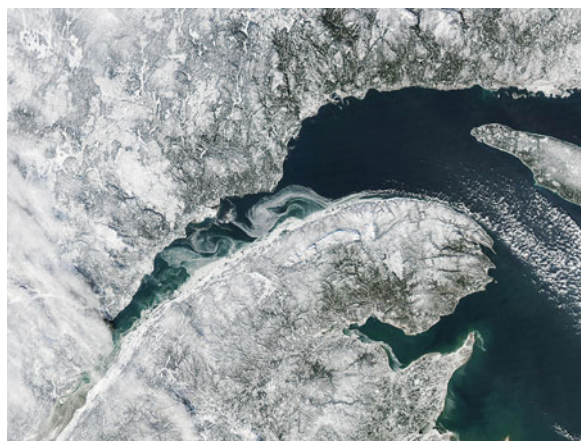
### Lake and River Ice

Lake and river ice form seasonally at mid and high latitudes and elevations. River ice forms under the flow of turbulent water which governs its thickness. The combination of ice jams on rivers with increased water flow during springtime snowmelt can result in catastrophic floods. Lake ice forms under less dynamic conditions resulting in a smoother ice surface that acts as an insulator to the underlying water. Hence lacustrine biology is strongly influenced by the formation of the ice canopy. The start of ice formation and the start of springtime ice break up are proxy indicators for changes in local climate as well as impacts on the ability to navigate these waterways.

Streams, rivers, lakes of all sizes dot the landscape. Locally, ice cover observations can be made from aircraft. Regionally or globally, river and lake ice monitoring is challenging because physical dimensions (long but narrow rivers) often require high resolution instruments like medium- to high-resolution optical data (Fig. 11) or SAR to resolve details [89]. Moreover, since the exact date of key processes, such as the onset of ice formation or river-ice breakup are unknown, voluminous data sets are required to support large-scale studies.

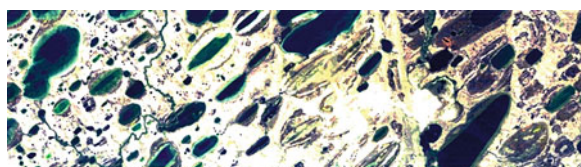
### Permafrost

Permafrost presents one of the greatest challenges for regional remote sensing technologies [7, 90]. The near surface active layer, the shallow zone where seasonal temperature swings allow for annual freeze and thaw, is complicated by different soil types and vegetative cover. This combination tends to hide the underlying persistently frozen ground from the usual airborne and space-borne techniques mentioned above. Even in winter when the active layer is frozen, remote sensing of the permafrost at depth is extremely difficult because of the spatially variable electrical properties of the material. Thus far, the most successful airborne and space-borne remote sensing methods involve optical photography



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 11**

January 2010 NASA MODIS image of ice formed on the St. Lawrence River. Thin layers of new ice are distorted into swirls by the surface currents (*center left*). Thicker ice is held fast to the southern shore (Image courtesy of MODIS Rapid Response Team, NASA GSFC)



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 12**

EO-1 Hyperion satellite data acquired over the north coast of Alaska (70.5 N 156.5 W) and about 50 km inland from the coast, which is toward the *right* on this image. The image is 7.5 km wide (*top to bottom*). Color composite using channels 16 23 and 28

to identify surface morphologies as proxy indicators of the presence of permafrost. Patterned ground and pingos are examples of the types of features visible in optical imagery and that are diagnostic of underlying permafrost. SAR interferometry has been suggested as another tool that can be used to monitor terrain for slumping associated with thawing permafrost. Figure 12 shows E01 Hyperion satellite, visible-band data collected over the north slope of Alaska. Shallow, oval shaped lakes form in thermokarst, which develops

when ice rich permafrost thaws and forms a hummocky terrain. Lakes and depressions left by drained lakes are densely distributed across the tundra. The long axis of the lake is oriented perpendicular to the prevailing wind direction. SAR intensity images have been used to determine that most of these small lakes freeze completely to the bottom during the winter months [91].

### Recent Developments in Airborne Radar Ice Sounding of Glaciers

Today, airborne radars operating between about 5 and 500 MHz are the primary tools used for measuring ice sheet thickness, basal topography, and inferring basal properties over large areas. These radars are typically operated as altimeters and acquire profile data only along nadir tracks that are often separated by 5 or more kilometers. The along track resolution is met by forming a synthetic aperture and the vertical resolution of the thickness is met by transmitting high bandwidth signal. Even though highly accurate thickness measurements can be achieved, information in the third cross-track dimension is absent.

While the surface properties of the ice sheets are becoming increasingly well documented, the nature of the glacier bed remains obscured by its icy cover. Revealing basal properties, such as the topography and the presence or absence of subglacial water, is important if we are to better estimate the flux of ice from the interior ice sheet to the sea and to forecast anticipated changes of the size of the ice sheets. Recent experiments demonstrate how it is possible to go beyond airborne nadir sounding of glaciers and to produce three-dimensional images of the glacier bed. This development represents a major step forward in ice sheet glaciology by providing new information about the basal boundary conditions modulating the flow of the ice and revealing for the first time geomorphologic processes occurring at the bed of modern ice sheets. The approach relies on the application of radar tomography to UHF/VHF airborne radar data collected using multiple, independent antennas and receivers. Tomography utilizes phase and amplitude information from the independent receivers to isolate the direction of a natural target relative to the aircraft. Combined with the range to the target based on the echo

travel time and position of the aircraft, tomographic methods yield swaths of reflectivity and topographic information on each side of the aircraft [92].

### Cooperative Efforts to Observe, Monitor, and Understand the Cryosphere

Several cooperative, international scientific studies of the high latitudes have been organized. Beginning with the 1897 voyage of the Belgica to the Antarctic Peninsula and continuing to the 2007 International Polar Year, studies have relied on the most recent technologies to increase knowledge of the polar regions. To realize the benefit of the growing constellation of international satellites to the scientific objectives of the 2007 International Polar Year (IPY) (Fig. 13), the Global Interagency IPY Polar Snapshot Year (GIIPSY) project engaged the science community to develop consensus polar science requirements and objectives that could best and perhaps only be met using the international constellation of earth observing satellites [93]. Requirements focused on all aspects of the cryosphere and range from sea ice to permafrost to snow cover and ice sheets. Individual topics include development of high-resolution digital elevation models of outlet glaciers using stereo optical systems, measurements of ice surface velocity using interferometric synthetic aperture radar, and frequently repeated measurements of sea ice motion using medium resolution optical and microwave imaging instruments.

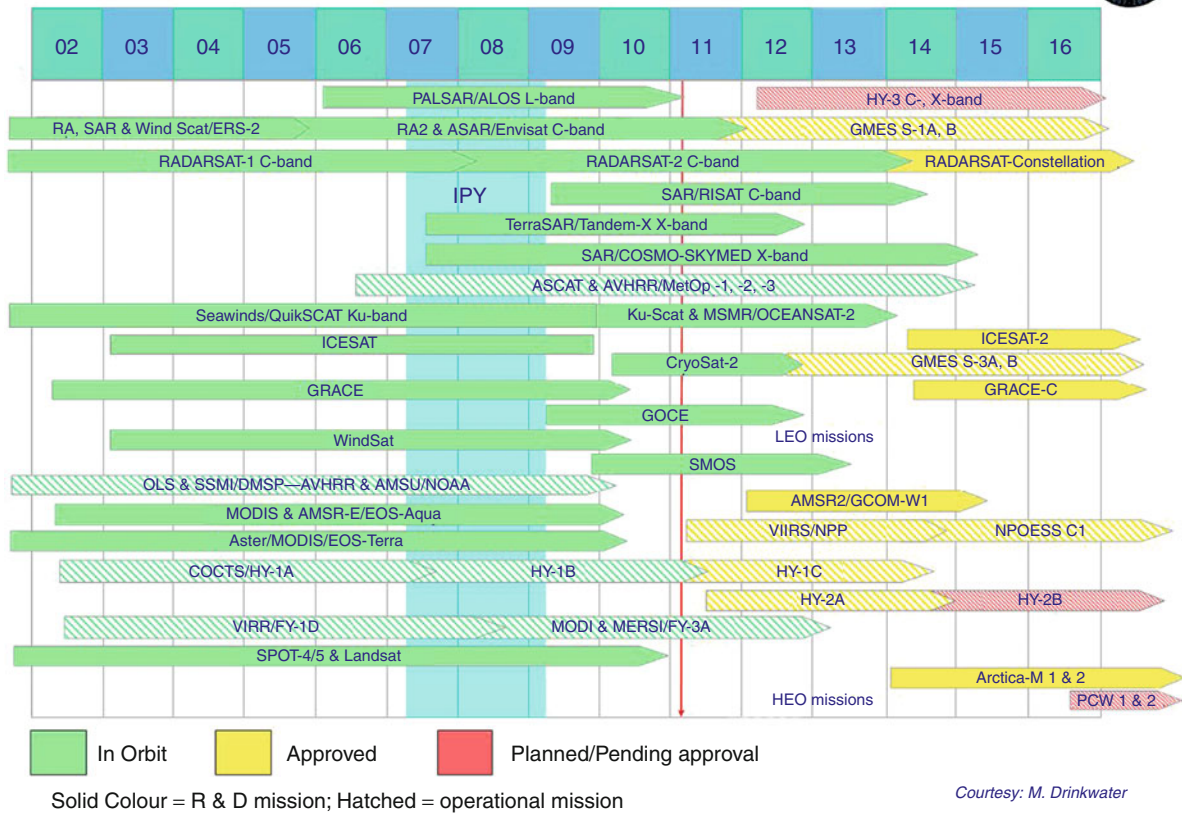
The IPY Space Task Group (STG), convened by the World Meteorological Organization (WMO), formed the functional link between the GIIPSY science community and the international space agencies. STG membership included representatives from the national space agencies of Italy, Germany, France, UK, US, Canada, Russia, China, Japan, and the European Space Agency (ESA), which in itself represents 19 nations. The STG determined how best to satisfy GIIPSY science requirements in a fashion that distributed the acquisition burden across the space agencies and recognized the operational mandates that guide the activities of each agency.

The STG adopted four primary data acquisition objectives for its contribution to the IPY. These are:

- Pole to coast multi-frequency InSAR measurements of ice-sheet surface velocity



### Cryosphere Satellite Missions



**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 13**

Current, approved and planned satellites for studying and monitoring the Cryosphere (Figure courtesy of Mark Drinkwater, European Space Agency)

- Repeat fine-resolution SAR mapping of the entire Southern Ocean sea ice cover for sea ice motion
- One complete high resolution visible and thermal IR (Vis/IR) snapshot of circumpolar permafrost
- Pan-Arctic high and moderate resolution Vis/IR snapshots of freshwater (lake and river) freeze-up and breakup

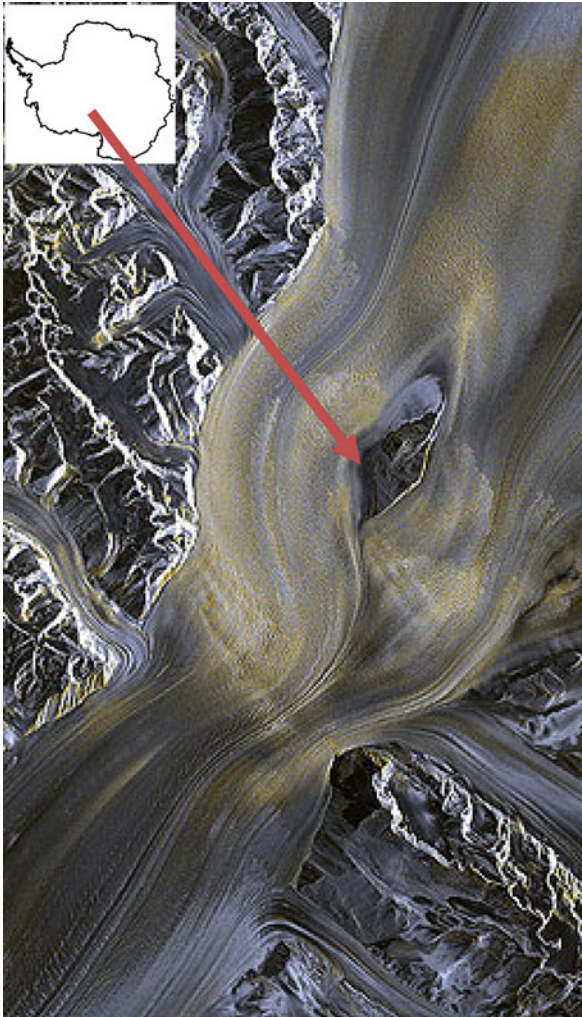
The STG achieved most of these objectives including: acquiring Japanese, ALOS L-band, ESA Envisat and Canadian Radarsat C-band, and German TerraSAR-X (Fig. 14) and Italian COSMO\_SKYMED X band SAR imagery over the polar ice sheets [94]; acquiring pole to coast InSAR data for ice sheet surface velocity; optically derived, high-resolution digital elevation models of the perimeter regions of ice caps

and ice sheets; coordinated campaigns to fill gaps in Arctic and Antarctic sea ice cover; extensive acquisitions of optical imagery of permafrost terrain; observations of atmospheric chemistry using the Sciamachy instrument.

### Future Directions

Emerging sensor and platform technologies hold great promise for future airborne and space-borne remote sensing of the cryosphere. Airborne programs are likely to continue to use large manned aircraft in a fashion similar to NASA's IceBridge program which integrates a sophisticated suite of instruments including lidars, radars, gravimeters, magnetometers, optical mapping systems, and GPS. But there will also be





**Airborne and Space-borne Remote Sensing of Cryosphere. Figure 14**

German Aerospace Center TerraSAR-X observations of the Nimrod Glacier (inset map of Antarctica). Ice floes around a central nunatak and down toward the Ross Ice Shelf. Crevasses appear in conjunction with the interruption of flow by the nunatak. Cooperative use of Canadian, German, European Space Agency, Italian, and Japanese synthetic aperture radar (SAR) satellites along with ground segment and data processing capabilities provided by the United States yielded a rich and diverse SAR data set that will be a lasting legacy of the IPY (TerraSAR-X data courtesy of D. Floricioiu, German Aerospace Center. Inset coastline derived from RADARSAT-1 Antarctic Mapping Project map)

a steady shift toward smaller, more dedicated unmanned aerial vehicles capable of remaining on station for longer periods and fulfilling some of the temporal coverage requirements that are difficult to satisfy with larger aircraft requiring a substantial number of crewmen.

It is worth noting again that ice sounding radars are exclusively deployed on aircraft for terrestrial research. This is driven operationally by the challenges of ionospheric distortions and by governmental controls on frequency allocations available for remote sensing applications designed to limit interference with communications and other commercial uses of the frequency bands. Certainly the latter is not an issue for other planetary studies and in fact the Martian ice caps have been successfully sounded from the orbiting MARSIS and SHARAD radars [95]. These extraterrestrial successes motivate continuing interest in deploying similar instrument for observing Earth's ice cover in the future.

As indicated in Fig. 13, there will be an ongoing constellation of satellites capable of collecting valuable cryospheric data. Here, coordination amongst the different space faring nations will be key to realizing the greatest scientific benefit. Lessons about cooperation gleaned from the IPY can be profitably extended to the acquisition of data and the development of geophysical products beyond the polar regions to all sectors of the cryosphere. There could also be generally better integration of the atmospheric chemistry and polar meteorological communities, as well as incorporation of gravity and magnetic geopotential missions into the coordination discussions. It is also possible to envision discussion and collaboration on emerging technologies and capabilities such as the Russian Arktika Project [96] and advanced subsurface imaging radars. A primary objective of continued coordination of international efforts is securing collections of space-borne "snapshots" of the cryosphere through the further development of a virtual Polar Satellite Constellation [97]. A natural vehicle for adopting lessons learned from GIIPSY/STG into a more encompassing international effort could be the Global Cryosphere Watch [98] recently proposed by WMO to be in support of the cryospheric science goals specified for the Integrated Global Observing Strategy Cryosphere Theme [1].

## Bibliography

### Primary Literature

- IGOS (2007) Integrated global observing strategy cryosphere theme report – for the monitoring of our environment from space and from Earth. WMO/TD-No. 1405. World Meteorological Organization, Geneva, 100p
- Perovich D, Light B, Eicken H, Jones K, Runciman K, Nghiem S (2007) Increasing solar heating of the Arctic ocean and adjacent seas, 1979–2005: attribution and role in the ice-albedo feedback. *Geophys Res Lett* 34:L19505. doi:10.1029/2007GL031480
- Perovich D, Richter-Menge J, Jones K, Light B (2008) Sunlight, water and ice: extreme arctic sea ice melt during the summer of 2007. *Geophys Res Lett* 35:L11501. doi:10.1029/2008GL034007
- Ainley D, Tynan C, Stirling I (2003) Sea ice: a critical habitat for polar marine mammals. In: Thomas D, Dieckmann G (eds) *Sea ice: an introduction to its physics, chemistry, biology and geology*. Blackwell Science, Oxford, pp 240–266
- Arrigo K (2003) Primary production in sea ice. In: Thomas D, Dieckmann G (eds) *Sea ice: an introduction to its physics, chemistry, biology and geology*. Blackwell Science, Oxford, pp 143–183
- Lizotte M (2003) The microbiology of sea ice. In: Thomas D, Dieckmann G (eds) *Sea ice: an introduction to its physics, chemistry, biology and geology*. Blackwell Science, Oxford, pp 184–210
- Grosse G, Romanovsky V, Jorgenson T, Water Anthony K, Brown J, Overduin P (2011) Vulnerability and feedbacks of permafrost to climate change. *EOS* 92(9):73–74
- Cazenave A, Llovel W (2010) Contemporary sea level rise. *Ann Rev Mar Sci* 2:145–173
- Prowse TD, Bonsal B, Duguay C, Hessen D, Vuglinsky V (2007) River and lake ice. In: Eamer J, Ahlenius H, Prestrud P, United Nations Environment Programme et al (eds) *Global outlook for ice and snow*. United Nations Environment Programme, Nairobi, pp 201–214. ISBN 978-92-807-2799-9
- Kim Y, Kimball J, McDonald K, Glassy J (2011) Developing a global data record of daily landscape freeze/thaw status using satellite passive microwave remote sensing. *IEEE Trans Geosci Remote Sens* 49(3):949–960
- McKinley AC (1929) *Applied aerial photography*. Wiley, New York, 341p
- Mittelholzer W and others (1925) *By airplane towards the north pole* (trans: Paul E, Paul C). Houghton Mifflin, Boston, 176p
- Wilkins H (1929) The Wilkins-Hearst Antarctic Expedition, 1928–1929. *Geogr Rev* 19(3):353–376
- Wilkins H (1930) Further Antarctic explorations. *Geogr Rev* 20(3):357–388
- Byrd RE (1930) *Little America*. G.P. Putman's Sons, New York, 422p
- McDonald RA (1995) Corona: success for space reconnaissance, a look into the cold war and a revolution in intelligence. *Photogramm Eng Remote Sens* 61(6):689–720
- Peebles C (1997) *The Corona Project: America's First Spy Satellites*. Naval Institute Press, Annapolis, 351p
- Wheelon AD (1997) Corona: the first reconnaissance satellites. *Phys Today* 50(2):24–30
- Richelson JT (1998) Scientists in black. *Sci Am* 278(2):48–55
- Sohn HS, Jezek KC, van der Veen CJ (1998) Jakobshavn Glacier, West Greenland: 30 years of spaceborne observations. *Geophys Res Lett* 25(14):2699–2702
- Zhou G, Jezek KC (2002) 1960s era satellite photograph mosaics of Greenland. *Int J Remote Sens* 23(6):1143–1160
- Bindshadler RA, Vornberger P (1998) Changes in the West Antarctic ice sheet since 1963 from declassified satellite photography. *Science* 279:689–692
- Kim K, Jezek KC, Sohn H (2001) Ice shelf advance and retreat rates along the coast of Queen Maud Land, Antarctica. *J Geophys Res* 106(C4):7097–7106
- Kim K, Jezek K, Liu H (2007) Orthorectified image mosaic of the Antarctic coast compiled from 1963 Argon satellite photography. *Int J Remote Sens* 28(23–24):5357–5373
- Waite AH, Schmidt SJ (1962) Gross errors in height indication from pulsed radar altimeters operating over thick ice or snow. *Proc IRE* 50(6):1515–1520
- Bogorodsky VV, Bentley CR, Gudmandsen PE (1985) *Radioglaciology*. D. Reidel, Dordrecht, 254p
- Fisher E, McMechan G, Gorman M, Cooper A, Aiken C, Ander M, Zumbege M (1989) Determination of bedrock topography beneath the Greenland ice sheet by three-dimensional imaging of radar sounding data. *J Geophys Res* 94(B3):2874–2882
- Koenig L, Martin S, Studinger M (2010) Polar airborne observations fill gap in satellite data. *EOS* 91(38):333–334
- Schanda E (1986) *Physical fundamentals of remote sensing*. Springer, Berlin, 187p
- Hall D, Martinec J (1985) *Remote sensing of ice and snow*. Chapman and Hall, New York, 189p
- Rees WG (2006) *Remote sensing of snow and ice*. Taylor and Francis Group, Boca Raton, 285p
- Petrenko VF, Whitworth RW (1999) *Physics of ice*. Oxford University Press, Oxford, 373p
- Carsey FD (ed) (1992) *Microwave remote sensing of sea ice*, A. G.U. geophysical monograph 68. American Geophysical Union, Washington, DC, 462p
- Hollinger J, Peirce J, Poe G (1990) SSM/I instrument evaluation. *IEEE Trans Geosci Remote Sens* 28(5):781–790
- Parkinson C, Gloersen P (1993) Global sea ice cover. In: Gurney R, Foster J, Parkinson C (eds) *Atlas of satellite observations related to global change*. Cambridge University Press, Cambridge, pp 371–383
- Zwally HJ, Yi D, Kwok R, Zhao Y (2008) ICESat measurements of sea ice freeboard and estimates of sea ice thickness in the Weddell sea. *J Geophys Res* 113:C02S15. doi:10.1029/2007JC004284
- Kwok R, Cunningham G, Wensnahan M, Rigor I, Zwally HJ, Yi D (2009) Thinning and volume loss of the Arctic ocean sea ice cover: 2003–2008. *J Geophys Res* 114:C07005. doi:10.1029/2009JC005312



38. Williams R Jr, Hall D (1993) Glaciers. In: Gurney R, Foster J, Parkinson C (eds) *Atlas of satellite observations related to global change*. Cambridge University Press, Cambridge, pp 401–422
39. Swithinbank C (1973) Higher resolution satellite pictures. *Polar Rec* 16(104):739–751
40. Swithinbank C, Lucchitta BK (1986) Multispectral digital image mapping of Antarctic ice features. *Ann Glaciol* 8:159–163
41. U.S. Geological Survey (2010) *Satellite image atlas of glaciers of the world*. USGS Fact Sheet FS 2005-3056, 2p
42. Merson RH (1989) An AVHRR mosaic of Antarctica. *Int J Remote Sens* 10:669–674
43. Bindshadler R, Vornberger P (1990) AVHRR imagery reveals Antarctic ice dynamics. *EOS* 71:741–742
44. Ferrigno JG, Mullins JL, Stapleton JA, Chavez PS Jr, Velasco MG, Williams RS Jr (1996) *Satellite image map of Antarctica, Miscellaneous investigations map series 1-2560*. U.S Geological Survey, Reston
45. Fahnestock MR, Bindshadler RK, Jezek KC (1993) Greenland ice sheet surface properties and ice dynamics from ERS-1 SAR imagery. *Science* 262:1525–1530
46. Jezek KC (2008) The RADARSAT-1 Antarctic Mapping Project. BPRC Report No. 22. Byrd Polar Research Center, The Ohio State University, Columbus, 64p
47. Jezek KC (1999) Glaciologic properties of the Antarctic ice sheet from spaceborne synthetic aperture radar observations. *Ann Glaciol* 29:286–290
48. Jezek K (2003) Observing the Antarctic ice sheet using the RADARSAT-1 synthetic aperture radar. *Polar Geogr* 27(3): 197–209
49. Liu H, Jezek K (2004) A complete high-resolution coastline of Antarctica extracted from orthorectified Radarsat SAR imagery. *Photogramm Eng Remote Sens* 70(5):605–616
50. Scambos TA, Haran T, Fahnestock M, Painter T, Bohlander J (2007) MODIS-based mosaic of Antarctica (MOA) data sets: continent-wide surface morphology and snow grain size. *Remote Sens Environ* 111(2–3):242–257
51. Bindshadler R, Vornberger P, Fleming A, Fox A, Mullins J, Binnie D, Paulsen S, Granneman B, Gorodetzky D (2008) The Landsat image mosaic of Antarctica. *Remote Sens Environ* 112(12):4214–4226
52. Korona J, Berthier E, Bernarda M, Rémy F, Thouvenot E (2008) SPIRIT. SPOT 5 stereoscopic survey of Polar Ice: reference images and topographies during the fourth International Polar Year (2007–2009). *ISPRS J Photogramm Remote Sens* 64(2):204–212. doi:10.1016/j.isprsjprs.2008.10.005
53. Goldstein RM, Englehardt H, Kamb B, Frohlich R (1993) Satellite radar interferometry for monitoring ice sheet motion: application to an Antarctic ice stream. *Science* 262:1525–1530
54. Kwok R, Fahnestock M (1996) Ice sheet motion and topography from radar interferometry. *IEEE Trans Geosci Remote Sens* 34(1):189–199
55. Joughin I, Kwok R, Fahnestock M (1996) Estimation of ice-sheet motion using satellite radar interferometry: method and error analysis with application to Humboldt Glacier, Greenland. *J Glaciol* 42(142):564–575
56. Gray AL, Short N, Matter KE, Jezek KC (2001) Velocities and ice flux of the Filchner Ice Shelf and its tributaries determined from speckle tracking interferometry. *Can J Remote Sens* 27(3):193–206
57. Joughin I (2002) Ice-sheet velocity mapping: a combined interferometric and speckle-tracking approach. *Ann Glaciol* 34(1):195–201
58. Eldhuset P, Andersen S, Hauge EI, Weydahl D (2003) ERS tandem InSAR processing for DEM generation, glacier motion estimation and coherence analysis on Svalbard. *Int J Remote Sens* 24(7):1415–1437
59. Rignot E, Forster R, Isaacs B (1996) Mapping of glacial motion and surface topography of Hielo Patagonico Norte, Chile, using satellite SAR L-band interferometry data. *Ann Glaciol* 23:209–216
60. Surazakov A, Aizen V (2006) Estimating volume change of mountain glaciers using SRTM and map-based topographic data. *IEEE Trans Geosci Remote Sens* 44(10): 2991–2995
61. Joughin I, Gray L, Bindshadler R, Price S, Morse D, Hulba C, Mattar K, Werner C (1999) Tributaries of West Antarctic ice streams revealed by RADARSAT interferometer. *Science* 286:283–286
62. Stearns L, Jezek K, Van der Veen CJ (2005) Decadal scale variations in ice flow along Whillans ice stream and its tributaries, West Antarctica. *J Glaciol* 51(172): 147–157
63. Beem L, Jezek K, van der Veen CJ (2010) Basal melt rates beneath the Whillans ice stream, West Antarctica. *J Glaciol* 56(198):647–654
64. Luckman L, Queneyand D, Beven S (2007) The potential of satellite radar interferometry and feature tracking for monitoring flow rates of Himalayan glaciers. *Remote Sens Environ* 111:172–181
65. Floricioiu D, Eineder M, Rott H, Yague-Martinez N, Nagler T (2009) Surface velocity and variations of outlet glaciers of the Patagonia Icefields by means of TerraSAR-X. In: *Geoscience and remote sensing symposium, IGARSS 2009*, vol 2, Cape Town, 12–17 Jul 2009, pp 1028–1031
66. Forster R, Rignot E, Isaacs B, Jezek K (1999) Interferometric radar observations of the Hielo Patagonico Sur, Chile. *J Glaciol* 45(150):325–337
67. Paul F, Haeberli W (2008) Spatial variability of glacier elevation changes in the Swiss Alps obtained from two digital elevation models. *Geophys Res Lett* 35:L21502. doi:10.1029/2008GL034718
68. Yu J, Liu H, Jezek K, Warner R, Wen J (2010) Analysis of velocity field, mass balance, and basal melt of the Lambert Glacier system by incorporating Radarsat SAR interferometry and ICESat laser altimeter measurements. *J Geophys Res* 115: B11102. doi:10.1029/2010JB007456
69. Thomas RH (1993) Ice sheets. In: Gurney R, Foster J, Parkinson C (eds) *Atlas of satellite observations related to global change*. Cambridge University Press, Cambridge, pp 385–400

70. Bhattacharya I, Jezek K, Wang L, Liu H (2009) Surface melt area variability of the Greenland ice sheet: 1979–2008. *Geophys Res Lett* 36:L20502. doi:10.1029/2009GL039798
71. Liu H, Wang L, Jezek K (2006) Spatio-temporal variations of snow melt zones in Antarctic ice sheet derived from satellite SMMR and SSM/I data (1978–2004). *J Geophys Res* 111: F01003. doi:10.1029/2005JF0000318
72. Rignot E, Thomas R (2002) Mass balance of the polar ice sheets. *Science* 297(5586):1502–1506
73. Rignot E, Kanagaratnam P (2006) Changes in the velocity structure of the Greenland ice sheet. *Science* 311(5673): 986–990
74. Zwally HJ, Giovinetto M, Li J, Cornejo H, Beckley M, Brenner A, Saba J, Yi D (2005) Mass changes of the Greenland and Antarctic ice sheets and shelves and contributions to sea-level rise: 1992–2002. *J Glaciol* 51(175):509–527
75. Wingham DJ, Shepherd A, Muir A, Marshall G (2006) Mass balance of the Antarctic ice sheet. *Philos Trans R Soc A* 364:1627–1635
76. Larsen CF, Motyka RJ, Arendt AA, Echelmeyer KA, Geissler PE (2007) Glacier changes in southeast Alaska and northwest British Columbia and contribution to sea level rise. *J Geophys Res Earth* 112:F01007
77. Thomas R, Frederick E, Krabill W, Manizade S, Martin C (2006) Progressive increase in ice loss from Greenland. *Geophys Res Lett* 33:L10503. doi:10.1029/2006GL026075
78. Herzfeld UC, McBride PJ, Zwally HJ, Dimarzio J (2008) Elevation change in Pine Island Glacier, Walgreen Coast Antarctica, based on GLAS (2003) and ERS-1(1995) altimeter data analyses and glaciological implications. *Int J Remote Sens* 29(19): 5533–5553. doi:10.1080/01431160802020510
79. Chen J, Wilson C, Blankenship D, Tapley B (2009) Accelerated Antarctic ice loss from satellite gravity measurements. *Nat Geosci* 2. doi:10.1038/NNGEO694
80. Luthcke SB, Zwally HJ, Abdalati W, Rowlands D, Ray R, Nerem R, Lemoine F, McCarthy J, Chinn D (2006) Recent Greenland ice mass loss by drainage system from satellite gravity observations. *Science* 314(5803):1286–1289
81. Velicogna I (2009) Increasing rates of ice mass loss from the Greenland and Antarctic ice sheets revealed by GRACE. *Geophys Res Lett* 36:L19503. doi:10.1029/2009GL040222
82. Luthcke S, Arendt A, Rowlands D, McCarthy J, Larsen C (2008) Recent glacier mass changes in the Gulf of Alaska region from GRACE mascon solutions. *J Glaciol* 54(188):767–777
83. Thomas R, Davis C, Frederick E, Krabill W, Li Y, Manizade S, Martin C (2008) A comparison of Greenland ice-sheet volume changes derived from altimetry measurements. *J Glaciol* 54(185):203–212
84. Rignot E, Velicogna I, van den Broeke MR, Monaghan A, Lenaerts J (2011) Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise. *Geophys Res Lett* 38:L05503. doi:10.1029/2011GL046583
85. Hall D, Riggs G, Salomonson V, DiGirolamo N, Bayr K (2002) MODIS snow-cover products. *Remote Sens Environ* 83: 181–194
86. Dery S, Brown R (2007) Recent northern hemisphere snow cover extent trends and implications for the snow-albedo feedback. *Geophys Res Lett* 34:L22504. doi:10.1029/2007GL031474
87. Foster J, Chang A (1993) Snow cover. In: Gurney R, Foster J, Parkinson C (eds) *Atlas of satellite observations related to global change*. Cambridge University Press, Cambridge, pp 361–370
88. Forster R, Long D, Jezek K, Drobot S, Anderson M (2001) The onset of Arctic sea-ice snow melt as detected with passive and active microwave remote sensing. *Ann Glaciol* 33:85–93
89. Jeffries M, Morris K, Kozlenko N (2005) Ice characteristics and processes, and remote sensing of frozen rivers and lakes. In: Duguay C, Piertroniro A (eds) *Remote sensing of northern hydrology*, Geophysical monograph series 163. American Geophysical Union, Washington, DC, pp 63–90
90. Duguay C, Zhang T, Leverington D, Romanovsky V (2005) Satellite remote sensing of permafrost and seasonally frozen ground. In: Duguay C, Piertroniro A (eds) *Remote sensing of northern hydrology*, Geophysical monograph series 163. American Geophysical Union, Washington, DC, pp 91–142
91. Jeffries M, Morris K, Liston G (1996) Method to determine lake depth and water availability on the north slope of Alaska with spaceborne imaging radar and numerical ice growth modeling. *Arctic* 49(4):367–374
92. Jezek K, Wu X, Gogineni P, Rodriguez E, Freeman A, Fernando-Morales F, Clark C (2011) Radar images of the bed of the Greenland ice sheet. *Geophys Res Lett* 38:L01501. doi:10.1029/2010GL045519
93. Jezek K, Drinkwater M (2010) Satellite observations from the International Polar Year. *EOS Trans AGU* 91(14):125–126
94. Crevier Y, Rigby G, Werle D, Jezek K, Ball D (2010) A RADARSAT-2 snapshot of Antarctica during the 2007–08 IPY. *News! Can Antarct Res Netw* 28:1–5
95. Picardi G, Plaut JJ, Biccari D, Bombaci O, Calabrese D, Cartacci M, Cicchetti A, Clifford SM, Edenhofer P, Farrell WM, Federico C, Frigeri A, Gurnett DA, Hagfors T, Heggy E, Herique A, Huff RL, Ivanov AB, Johnson WTK, Jordan RL, Kirchner DL, Kofman W, Leuschen CJ, Nielsen E, Orosei R, Pettinelli E, Phillips RJ, Plettemeier D, Safaenili A, Seu R, Stofan ER, Vannaroni G, Watters TR, Zampolini E (2005) Radar soundings of subsurface Mars. *Science* 310(5756):1925–1928. doi:10.1126/science.1122165
96. Asmus VV, Dyaduchenko VN, Nosenko YI, Polishchuk GM, Selin VA (2007) A highly elliptical orbit space system for hydrometeorological monitoring of the Arctic region. *WMO Bull* 56(4):293–296
97. Drinkwater MR, Jezek KC, Key J (2008) Coordinated satellite observations during the International Polar Year: towards achieving a Polar Constellation. *Space Res Today* 171:6–17
98. Goodison B, Brown J, Jezek K, Key J, Prowse T, Snorrason A, Worby T (2007) State and fate of the polar cryosphere, including variability in the Arctic hydrologic cycle. *WMO Bull* 56(4): 284–292

## Books and Reviews

- Gloersen P, Campbell W, Cavalieri D, Comiso J, Parkinson C, Zwally H (1992) Arctic and Antarctic sea ice, 1978–1987: satellite passive microwave observations and analysis, NASA SP-511. NASA, Washington, DC, 290p
- Parkinson C, Comiso J, Zwally H, Cavalieri D, Gloersen P, Campbell W (1987) Arctic sea ice, 1973–1976: satellite passive microwave observations, NASA SP-489. NASA, Washington, DC, 296p
- Schnack-Schiel S (2003) The macrobiology of sea ice. In: Thomas D, Dieckmann G (eds) Sea ice: an introduction to its physics, chemistry, biology and geology. Blackwell Science, Oxford, pp 211–239
- Weeks W, Hibler W (2010) On sea ice. University of Alaska Press, Fairbanks, 664p
- Zwally H, Comiso J, Parkinson C, Campbell W, Carsey F, Gloersen P (1983) Antarctic sea ice, 1973–1976: satellite passive microwave observations, NASA SP-459. NASA, Washington, DC, 206p

## Airborne Toxic Chemicals

APRIL HISCOX<sup>1</sup>, MARK MACAUDA<sup>2</sup>

<sup>1</sup>Department of Geography, The University of South Carolina, Columbia, SC, USA

<sup>2</sup>Health Promotion, Education, and Behavior, University of South Carolina, Columbia, SC, USA

## Article Outline

Glossary  
 Definition of the Subject and Its Importance  
 Introduction  
 Toxics in the Air  
 Regulation and Monitoring  
 The Toxic Cycle: Specific Examples  
 Human Health  
 Future Directions  
 Bibliography

## Glossary

**Air toxic** Substances that are known or suspected to cause cancer or other serious health effects. Also known as Hazardous Air Pollutants (HAPS).

**Anthropogenic source** A source of air toxics created by human beings.

**Area source** A single source of pollutant that emits less than 10 t per year of one air toxic, or less than 25 t per year of any combination of air toxics.

## MACT (Maximum Achievable Control Technology)

Standard that dictates emission limits of a source is set by the best performing 12% of similar sources, if more than 30 similar sources exist nationally. If less than 30 exist, the best five are used to set the standard.

**Major source** A single source of pollutant that emits 10 t per year or more of one air toxic, or 25 t per year or more of any combination of air toxics.

**Mobile source** A source of air toxics that moves (such as a car, truck, airplane, or boat).

## Definition of the Subject and Its Importance

“Air toxics” is a term that is often used colloquially, but for the purposes of this article it will be used following the specific definition set forth by the US Environmental Protection Agency (EPA). Toxic (or Hazardous) air pollutants are those pollutants that are known or suspected of causing cancer or other serious health effects [1]. Air toxics are defined by the Clean Air Act, which explains what pollutants qualify, how different sources are categorized, and how they are to be regulated [2]. The Clean Air Act (CAA) is the most comprehensive modern legislation concerning air quality in the USA. The original 1970 Act defines “Hazardous Air Pollutants” (HAPS) or air toxics as those substances known or suspected of causing cancer, birth defects, or other adverse health problems. It charges the EPA to “Significantly reduce emissions of the most potent air pollutants” [2]. Initially, the Clean Air Act listed 189 chemicals as toxics; during the revision of 1990 caprolactam was eliminated based on new scientific evidence, leaving 188 [1].

## Introduction

It can be said that air pollution is as old as civilization itself. With the discovery of fire, humans started emitting substances into the atmosphere. Ancient Rome had issues with pollution of land, water, and air, including emissions from copper and lead production that were greater than Europe during the nineteenth century [3]. In 1954, smog descended on London, primarily from coal burning, for 4 days and killed 4,000 people [4–6]. Air pollution has been known to affect evolution of species; the most famous example being the peppered moths of England, who evolved a darker color to match soot covered trees near industrial operations [7].

Despite its long history and fundamental importance to environmental health and quality of life, air quality is one of the harder aspects of environmental health to grasp, and even harder to control. Unlike water or soil, air cannot be picked up and held. How it travels from place to place is still the subject of intense scientific research. Human understanding of the air, how pollutants enter into it and travel, and how they ultimately affect human and animal health is still being actively investigated. This chapter will summarize the fundamentals of air toxics, a subset of air pollutants, from both an environmental and human health perspective. The following presents a synopsis of the current state of knowledge and discusses future directions for research in this area. The article will start off with a brief summary of air toxics, where they come from and where they go. Readers are encouraged to consult the numerous air pollution textbooks listed in the references for an in-depth treatment of these topics. Following this background material, three specific toxics, benzene, mercury, and perchloroethylene and their roles in modern living will be discussed. The article will finish with a discussion of future directions for better understanding of the role of air toxics from a sustainability standpoint.

## Toxics in the Air

### Where Do They Come From?

Any introductory textbook on the topic of air pollution starts with the fundamental knowledge that regulated pollutants are from both biogenic and anthropogenic sources. This information is typically presented in the context of the six major criteria pollutants, which constitute the EPA regulations for ambient air quality. Toxics, however, represent a different set of pollutants and by definition it is not necessarily their ambient concentration that is of main concern. It is important to recognize that while some toxics can be emitted from natural occurrences such as volcanic eruptions and forest fires, by and large, toxics come from anthropogenic sources.

Air toxics can be emitted into the atmosphere in several ways. Attrition occurs through the mechanical wearing of physical objects. Activities such as grinding, polishing, sanding, drilling, and spraying can lead to pollution by attrition. This can be found in activities such as drilling for oil [8], or in a local auto body

shop that specializes in sanding and painting cars. Vaporization occurs when a liquid converts into its gaseous form. This can happen under temperature and pressure, or because the liquid is volatile (readily evaporates at normal temperatures). Gasoline is one such volatile liquid that readily evaporates [8]. One of the byproducts of gasoline evaporation is benzene, which is discussed in detail below. The third manner in which many toxics become airborne is through combustion. Combustion occurs when a substance is combined with oxygen in a chemical reaction that creates energy. If a fuel is perfectly combustible, meaning that all of it gets used in the combustion process, the two outputs are water vapor and carbon dioxide. Both of which, while not toxic to humans are greenhouse gasses that can contribute to global warming through the trapping of heat energy from the sun. Most fuels, however, are imperfectly combustible, meaning that they do not burn completely. This leaves byproducts such as benzene, toluene, and formaldehyde [8].

For a full understanding of the sources of air toxics and regulation, one must first be familiar with the source categories. The main concern for air toxics regulators are routine emissions: Those that are produced as a byproduct of a process, rather than accidental “one time” emissions [1]. Three major source categories exist. *Mobile sources* include any mechanical object that moves (cars, planes, trains, marine, farming equipment, etc.). Over 50% of air toxic released into the atmosphere come from mobile source emissions [9]. In addition to emitting greenhouse gasses, auto emissions also include several air toxics such as formaldehyde, acetaldehyde, 1,3-butadiene, and particulate matter from diesel engines. These are all considered possible carcinogens by the EPA. The EPA estimates that over half of all cancers caused by outdoor emissions are caused by motor vehicle emissions [9].

While motor vehicles contribute about half of hazardous air toxics, air toxics are also released by stationary sources, which are those sources that do not move. Stationary sources can fall into one of two categories: *Major sources* are a single (or point) source that emits 10 t per year of any one listed toxic air pollutant or 25 t of a mixture of any listed air toxic pollutants. *Area sources* are those single sources that emit less than 10 t of one, or 25 t of more than one air pollutant.

Area sources may not contribute a large portion of pollution on their own, but coupled with other small sources, such as one might find in a city, area sources can significantly impact air quality. As the population becomes denser, the impact of these small sources on overall air quality becomes greater; since there are more sources of toxics per unit area (such as combustion sources from cooking and heating). The urban poor are especially affected since they tend to live in more highly populated areas, as is the case in Hong Kong, where those more financially well off can live farther from the populous city center [10].

It should be noted that though sources that emit toxics into the atmosphere are of primary concern, a HAP may pollute an indoor area as well, such as coal burning for cooking or heating. Indoor sources present a concern as they can be contained in smaller spaces, and therefore expose those occupying that space (such as industry workers) to high concentrations of toxics. Indoor sources pose high levels of risk in developing areas like parts of China, where coal is used for home heating [11]. The full list of HAPs contains a wide array of compounds ranging from industrial chemicals to agricultural pesticides, which can be present in the air in particulate (solid), gaseous, or liquid (aerosol) form.

### Where Do They Go?

Once a pollutant becomes airborne, it can have many potential pathways. It can remain in the air as is, become a component of a chemical reaction and transform, and/or it can be transported short or long distances and then follow another pathway. The complexities of each of these mechanisms depend on numerous factors ranging from the scale of long-term climates to the scale of short-term turbulence occurring within seconds of release. As long as the toxic is in the air, it can affect the global climate or an individual human in a single breath. The complexities of atmospheric interactions of chemicals can complicate efforts to reduce toxics; Each chemical that is emitted into the atmosphere can react differently with other chemicals, to the point that the reduction of one type of pollutant (such as  $\text{NO}_x$ ) can actually increase the production of other pollutants (such as Ozone and aldehydes), depending on solar radiation, climate, and season [12].

The fate of pollutants can be affected by a number of complex physical transformations, including nucleation and coagulation (by which particles grow in size) or deposition through settling or precipitation. Toxics that leave the air and are deposited on the ground can continue to do environmental and health harm by contaminating land and water (including drinking water). Humans can then be exposed by drinking contaminated water, eating contaminated food, or by coming in contact with contaminated soils [8]. Toxics that have settled or precipitated out can also find their way into the food chain. As larger animals eat smaller animals that are contaminated, toxins can accumulate in biologic tissues, resulting in greater concentrations as one rises higher in the food chain. Human consumption of animals higher in the food chain can result in significant doses of toxin. The classic example of this is mercury contamination of tuna fish, which accumulate the toxin through feeding on smaller aquatic life forms [13].

The interactions between air toxics and the atmosphere is multilayered and dynamic, factors such as time of day, season, wind directions, and temperature can affect what happens to a toxic after it is released into the air. Concentrations of mercury, for example, increase in Korea during winter and at night because of increased coal burning. In contrast, in some areas of China, gaseous mercury concentrations increase in summer because high solar radiation transforms mercury trapped in soil into a gaseous form [11]. Some studies have shown that air toxics tend to have periods of low ambient concentrations that are followed by sharp spikes in output. Logue, Huff-Hartz et al. found that over 50% of the toxics measured by their high time resolved methods occurred during spikes in emissions of short duration; and that different areas (such as high traffic areas, areas next to industrial sites, and general city buildings) can have very different release profiles, even if they are in close proximity (in this case 13 km) from one another [14]. Further examples of the importance of microclimates include Raymer et al. who showed that exposures to aldehydes (which are toxics formed from combustion engines, cigarette smoking, oil frying) have been shown to vary widely in different microenvironments, being higher in restaurants but lower in gas stations in the same city [15]. Similarly, areas of high automobile traffic have been shown to



equate with high levels of air toxics [16]. With timing and specific location being an important part of the exposure equation, individual human movement patterns are an important part of one's risk for exposure to air toxics [16]. Two individuals living in the same city, or even the same block, can have different exposures depending on when they are present in that block relative to the time of day that an exposure might be taking place.

Though two individuals living in the same city might be exposed to different levels of toxics, they are both likely to be exposed to worse air quality than their rural counterparts. Cities, where most anthropogenic sources are centered, present unique air quality challenges that go beyond the mere high concentration of toxic pollutants due to population. The physical layout of a city affects the air that moves within and around it. The buildings in large cities create a larger surface area to collect heat in the daytime, which they reradiate at night. Warm pollutant filled air concentrates in locations with high numbers of tall buildings (usually at a city's center). That air rises and spreads out over the city, cooling as it moves. As it reaches the city's edge it is drawn back in to fill the void left by the rising warm air in the center. The result is a convection current that circulates pollutants within the city [8]. Though particular microclimates within a city will have their own levels of air toxics based upon what is occurring locally [15], the overall air quality in a city can be worse than it is in less populated areas because of the properties of the physical landscape.

## Regulation and Monitoring

As one might imagine, addressing the potential health implications of almost 200 different compounds, all with different sources, means of transport, fates and health effects presents a complex problem. During attempts to regulate air toxics, the EPA has tried several strategies. From 1970 to 1990, the EPA attempted to set standards and regulate each of the 189 toxic air chemicals based on the individual health risks that were posed by each one. The strategy was to identify all of the pollutants that could cause "serious and irreversible illness and death" and reduce the emissions of each to a point that provided a margin of safety to the public. Issues arose with this approach however, as

the EPA attempted to create policy based on incomplete scientific evidence; pinpointing the level of reduction needed to avoid health effects proved to be easier said than done! How risk was to be assessed and the level of acceptable risk that should be incurred by the public for each of the pollutants created an inefficient and slow system that only saw regulation of seven pollutants in 20 years [1].

In 1990, with the revision of the Clean Air Act, congress charged the EPA to implement a new system of regulation using a technology-based approach called "Maximum Achievable Control Technology" or MACT. The MACT standard dictates that the emissions limits of a certain toxin be set by the average emissions of the best performing 12% of similar sources, if there are more than 30 sources nationally that are in the same category. If there are less than 30, the average emissions of the best five are used as a standard [1]. Since air toxics may still be harmful even at the emissions levels of the best emitters, the EPA is able to assess how well current technologies reduce risks and has the power to implement additional standards to deal with any remaining risk posed by generation of toxic pollutants. The EPA must explore the remaining health risks posed by a pollutant 8 years after issuing the MACT standard [1].

The 1990 amendments to the Clean Air Act also eliminated caprolactam as a HAP. This shows that the understanding of pollution and regulation is always changing and that the links between regulations, reductions of pollutants, and the health effects of pollutants are still being actively explored [17–20].

In order to evaluate toxic levels and create appropriate regulations, there must be an understanding of how much toxic is being emitted and how it behaves in the environment. Understanding how air toxics behave, the concentration level at any one time, and how and when an individual might be exposed is an evolving science. Both direct sampling of air quality and modeling of air toxics based upon preexisting data (or a combination of both) are used. The Environmental Protection Agency does not monitor the entire USA in order to gain an understanding of the concentration of air toxics; rather they receive reports from industries who account for their own emissions, which are compiled in the Toxic Release Inventory Report [21]. These emissions data are combined with data such as the rate

of toxic release, the location, the height of the release, the nature of the pollutants with respect to decay and longevity, wind speed, and wind direction. This information is then broken down by census tract and used in models such as the Assessment System for Population Exposure Nationwide (ASPEN) in order to estimate the level of pollutants in a given area of USA [18, 20, 22]. The success and accuracy of a model depends upon the assumptions made and the inputs used, and different modeling techniques can produce different results for the same area [16].

Direct sampling of toxics can also take different forms. Some techniques, for example, collect an average concentration over a period of time, such as 24 h. Others are able to sample at more regular intervals, allowing a profile of different concentrations during different times of the day [14]. In addition, there may be several different sampling technologies for one pollutant, such as is the case with mercury [11]. *When sampling occurs is also important:* Concentrations can vary both by season, and by time of day. Sampling can help to test the validity of models, and help to determine what factors can make a model more accurate [18, 23].

The difficulties in monitoring and sampling air toxics have led to recent controversy. In 2009, USA Today and scientists from University of Maryland and Johns Hopkins tested the air outside 95 schools nationwide and found that seven schools had high enough levels of toxics such as benzene and chromium to elevate risks of cancer, and 57 schools had levels that were higher than their respective state guidelines. The article created a public outcry, and in response to the article, Louisiana and Pennsylvania conducted their own short-term monitoring and found that levels were not high enough to pose a health threat. This led to questions about how areas are monitored, the duration of monitoring, and whether the health threats existed or not [24, 25].

Monitoring and Regulation of hazardous materials is complex. The ways in which air toxics behave in the atmosphere are not always completely understood, and monitoring and modeling air toxic behavior is subject to several variables that can lead to different results. This can lead to controversy about the level of exposure (and thus risk) of the general public. In order to regulate such a complex issue, the EPA has taken a technological approach, based upon the level of

control possible, which allows them to establish regulations despite incomplete knowledge about distribution and health effects.

## The Toxic Cycle: Specific Examples

### Mobile Sources: A Balancing Act

As mentioned above, mobile sources, most notably automobiles, represent a major contributor to air toxic emissions in the USA. Of high concern is benzene. Benzene is a colorless liquid with a sweet odor. It evaporates rapidly and is used in plastics, resins, and synthetic fibers. It also naturally occurs in crude oil and is present in gasoline when it is refined. The benzene that exists in gasoline can become airborne when gasoline evaporates or vaporizes. Most benzene, however, comes from incomplete combustion of other naturally occurring compounds in gasoline, namely, toluene and xylene [9]. Inhaling very high levels of benzene can cause death. At slightly lower levels, it can cause drowsiness, confusion, and increased heart rate. Most seriously, it has been linked to leukemia of the blood called AML, which is a byproduct of benzene's effect on blood cells and bone marrow [9, 26]. Children are of particular concern with respect to air toxics because their bodies are still growing, and air toxics can affect them developmentally [18, 20]. In a study by Whitworth et al., census tracts with the highest levels of air benzene based on ASPEN also had the highest levels of leukemia in children. Other studies have found that concentrations of benzene and 1,3-butadiene exceed EPA health benchmarks at hundreds of locations across the USA [18], making benzene a serious health threat.

In order to eliminate air toxics, including benzene, from gasoline, the 1990 revision of the Clean Air Act mandated that highly polluted cities use reformulated gasoline; which is required to be less likely to vaporize, and have lower levels of benzene and aromatics. In addition, in many states, gasoline pumps are required to have vapor recovery nozzles, in order to trap gasoline vapor that may evaporate during refueling [9]. Reformulation of gasoline to eliminate benzene has been successful. California saw reductions in benzene emissions of 54% in the mid-1990 s, corresponding with the introduction of reformulated gasoline [17].

Another requirement of reformulated gasoline is that it burns more efficiently. This last requirement is accomplished by adding compounds that oxygenate the gasoline, which allow gasoline to burn more completely and efficiently. Two such compounds are ethanol, and MTBE or methyl tertiary butyl ether. The goal of the increased efficiency is to reduce the production of carbon monoxide, a cause of smog and to conserve oil, which is a nonrenewable resource. Thus it serves several purposes [27].

Studies of the effectiveness and side effects of ethanol and MTBE gasoline have shown that the addition of these compounds does reduce CO emissions, but increases the concentrations of formaldehyde and ethyl-aldehyde [27]. Formaldehyde is one of the 188 EPA air toxics [28]. In addition, MTBE itself has been named as a health concern, causing irritation and nervous system effects [26]. Health concerns caused California to phase out MTBE in favor of ethanol as an oxygen booster [17]. The story of MTBE makes an important point about the efforts to lessen the impact of the industrial way of life on the environment, so that mankind sustains itself into the future. Interactions between chemicals and the atmosphere are complex, and adding something new to the equation can have unforeseen consequences, even if the original aim was to reduce the levels of a harmful substance. MTBE was introduced to help reduce smog, but it also increases levels of some air toxics, and has health concerns of its own.

The phase-in of reformulated gasoline highlights another issue in regulation; the economic, and by extension, political trade-offs inherent in regulating an industry: The phase-in of reformulated gasoline in California has led to higher gasoline prices for consumers, as well as disadvantaged smaller refiners who have had more trouble in meeting the regulations in relation to larger refiners [29]. Thus the regulation of air toxics can, and does, become a political and economic issue.

The most transformative solution to the problems of air toxics from gasoline is to simply greatly reduce or eliminate the use of gasoline altogether. In 2010, cars that run on both electric and gasoline power (i.e., hybrids) are gaining greater market share, and automotive manufacturers, both large and small, are designing and constructing vehicles that run

exclusively on electric power. A complete transition from gasoline to electric power creates its own obstacles, however. Battery technology is one of them. Traditional lead acid batteries are inexpensive, but are heavy and cannot hold enough power for electric vehicles to be practical. Nickel-metal hydride batteries have more power but can become permanently damaged if over discharged. Lithium ion batteries, a newer technology, can store a good deal of power and are not affected by discharge, but some use cobalt, which is highly toxic [30]. Thus the ideal battery technology is yet to be developed, and existing technologies have inherent trade-offs between efficiency, cost, and environmental impact.

Technology development is not the only obstacle to changing the automotive landscape: The use of Electric cars could be a benefit to energy companies, allowing them to gain revenue as cars charge at night (when little electricity is being used elsewhere). The integration of electric vehicles, however, would necessitate a fundamental change in power infrastructure, which would take time and serious amounts of coordination between industries, such as the power generation industry and the automobile industry, two industries that have never worked together very closely [30]. This brings up a very important issue when talking about sustainability. The challenges and obstacles to sustainable practices are not just limited to developing technology, but are also problems of coordination, commitment and political will, and balancing the trade-offs inherent in changing from one system to another.

### Perchloroethylene: Keeping Things Clean

A common site in cities (and suburbs) in the USA is a local, small neighborhood dry cleaner. In fact, there are about 27,000 free standing dry cleaners within the USA [31]. Many people, when entering a dry cleaners, will notice a particular odor; both in the establishment, and on their newly cleaned clothes. That odor is Perchloroethylene or PERC. It historically has been the most common chemical associated with dry cleaning. PERC has been linked to liver and kidney tumors in rats and is considered by the EPA to be a carcinogen. In high concentrations (such as that associated with occupational exposure) it is associated

with dizziness, confusion, nausea, difficulty in speaking, unconsciousness, and death [26]. Exposure to PERC can happen from inhaling the chemical directly or ingesting contaminated water. Once in the body, it can be stored in fat cells, and releases itself slowly into the blood stream [26].

The case of PERC provides a good example of how an air toxic is regulated, the different categories of emitters, and how technology provides the basis for guidelines. The 1990 revision of the Clean Air Act required regulation based upon available technology (MACT). The first rules for PERC dry cleaners were laid out in 1991, revised and finalized in 1993 and updated in 2005–2006 as technology improved and health effects were better understood. In regulating the industry, the EPA divided dry cleaners into three groups. The first were major sources (those that produce more than 10 t of PERC per year), the second were smaller area sources that were free standing or in shopping plazas, the third (which was a category that was not originally differentiated in the 1993 guidelines, but believed to cause a greater health threat) were dry cleaners that exist as part of residential buildings [31].

In 2005, the EPA estimated that nine million people lived within 6 miles of a major source of PERC [32]. The main strategy for the large cleaners (the major sources) is the requirement to use state-of-the-art recovery systems that do not permit PERC to be released into the air, but rather trap and filter it. They are also required to implement advanced leak detection practices to ensure that systems do not leak PERC into the atmosphere accidentally [31].

The guidelines for smaller area sources are based upon the standard of “Generally Available Control Technologies” or GACT. GACT is slightly less stringent than MACT and mandates that firms that emit HAPs use control technologies that are commercially available and the most appropriate considering the economic impact on regulated entities and the ability for entities to comply [33].

As per the GACT standard, new free-standing operations are required to build facilities with closed loop systems (that do not release PERC into the air) and implement advanced leak detection practices. Dry cleaners that already exist have to eliminate what are called transfer machines, those that require PERC soaked clothes to be moved to a second machine

for drying, in favor of machines that wash and dry together [32].

Dry cleaning operations that are part of apartment buildings pose a greater risk to the public because people are living in close proximity to PERC emissions; thus more stringent regulations were proposed. According to the rule, all existing dry cleaners have to eliminate transfer machines, and are not able to replace PERC machines when the old ones wear out. PERC machines will have to be phased out completely by 2020 [31].

These regulations are not without costs. For example, implementation of these regulations has been estimated to cost the smaller dry cleaners 7.3 million dollars collectively. Though greater efficiency might save 2.7 million annually, these changes represent an overall net cost to the industry [32].

Dry cleaning of clothes is a service millions of Americans use without giving much thought to the health effects or environmental impacts of the service. Even though a small neighborhood dry cleaner might not cause a large environmental problem on its own; a densely populated area, such as a city, might have tens or hundreds of small establishments. Taken together, they can have a large environmental impact. The regulations are foremost to protect health, and those operations that cause a greater immediate threat (such as residential dry cleaners) are regulated more strictly. The regulations also attempt to take into account the ability of emitters to follow the regulations in a way that is not detrimental to their business, hence the GACT guidelines for smaller area emitters. Regulations are often a give and take.

### Mercury: A Global Toxin

Air toxics are truly a global problem, and a good example of that is the harm posed by human exposure to mercury. Mercury is an element and is naturally occurring in the environment. It can be emitted through geothermal and volcanic activity. The main concern for human health, however, comes from human-made (or anthropogenic) sources of mercury emissions. In adults, mercury exposure can damage the brain, heart, kidneys, lungs, and the immune system. Brain effects can lead to personality changes, shyness, tremors, and hearing and sight problems. A bigger risk comes for children. Fetuses and young children exposed to mercury can develop neurological and developmental

disorders that can lead to mental retardation, blindness, seizures, inability to speak, and lack of coordination. It can also cause damage to kidney, liver, and digestive systems [34]. Mercury can be passed from mother to infant in breast milk or from mother to fetus directly during gestation.

Mercury poisoning goes far back into human history: Mercury poisoning can result from exposure during metalworking and gold extraction, and there is evidence of increased mercury concentrations globally during times of high levels of gold mining activity, including the ancient Roman and Incan empires [35]. Small-scale gold working is still a major source of mercury pollution in places such as Africa and Asia [36, 37]. Mining and metal working are not the only industries that historically exposed one to mercury: The term “Mad as a Hatter” comes from mercury poisoning among hatters, who used the element in the creation of hat felt from animal pelts and suffered cognitive effects [38].

Currently, humans introduce mercury into their environment through metal mining, waste incineration, refining, and manufacturing. The largest source of mercury comes from the combustion of coal for electricity generation and heating. In fact 60% of global mercury production in 2000 was due to coal burning [39]. Mercury can exist in several forms: Elemental mercury ( $\text{Hg}^0$ ), divalent mercury ( $\text{Hg}^{2+}$ ), and particulate mercury ( $\text{Hg}(\text{p})$ ). Elemental mercury is the most common form. It can exist in the atmosphere for extended periods of time, a year or more, which means that it can travel large distances, up to 1,000 km from its source. In fact, it has been argued that elemental mercury production in Asia contributes to mercury levels in North America [36]. Mercury occurs naturally in coal. The amount of mercury depends upon the type of coal and its origin. When coal is burned for fuel, elemental mercury, particulate mercury, and divalent mercury can be released. Elemental mercury can oxidize into  $\text{Hg}^{2+}$  and deposit in waterways and land.  $\text{Hg}^{2+}$  that is deposited in waterways can become transformed by certain bacteria into methyl mercury ( $\text{CH}_3\text{Hg}^+$ ). Though these basics are understood, the details about how mercury cycles through the environment are still under investigation [40].

Methyl mercury is highly toxic to humans and animals, and can accumulate in animal tissue. Through

this accumulation (called bioaccumulation) humans can receive large doses of mercury from ingesting several smaller animals that have been contaminated. Consumption of toxic fish is the primary risk of mercury globally. In the USA alone, 48 out of the 50 states had to advise residents to avoid consuming fish from certain bodies of water within their borders [39].

Mercury from coal-fired power plants is a particular problem since rapid development, and the associated need for power generation, has increased coal usage worldwide. Poland, for example, derives 93% of its power from coal and has no legal limitations on mercury emissions, making it the fourth highest emitter in Europe [41]. Asia, because of rapid development, is a source of mercury emissions from all sources, with a large portion coming from manufacturing [36] and coal burning [40].

The primary post-combustion technologies that reduce mercury emissions are not specifically for mercury control, but rather for controls of particulate matter,  $\text{SO}_2$  and  $\text{NO}_x$  [40, 42]. Combinations of control technologies can be successful for mercury reduction. South Korea, for example, has seen a 68% reduction in mercury emissions by using a combination of technologies intended for control of  $\text{SO}_2$  and  $\text{NO}_x$  [43]. Though reductions based on current technologies are possible, the lack of specific controls indicates that the science of mercury control is lacking. Even developed countries, such as the USA, do not have a clear direction in dealing with this hazard.

The path of regulation of mercury in the USA has taken a few interesting turns. In the USA, power plants are the largest emitters of mercury, and the Clean Air Act revisions of 1990 charged EPA to study coal burning in the USA. They found that about a third of the mercury found in coal was being removed by current filters (the ones primarily for  $\text{SO}_2$  and  $\text{NO}_x$ ), while the other two thirds were being released into the atmosphere. The EPA was in the process of developing new standards for mercury emissions under MACT until 2004, when they changed course and proposed a cap and trade system instead. In the cap and trade system, the mercury emissions would be limited nationally, and power plants would possess credits for each unit of mercury produced. Plants that produced *less* mercury than permitted could then sell their excess units to another power plant. That plant could then produce



more. The cap and trade system has come under fire, however, because though it would reduce emission of mercury nationally, it could create a situation where mercury emissions are high locally, affecting residents of an area disproportionately. It was ultimately struck down by the USA court of appeals, who ruled that the EPA did not follow the proper procedures for delisting mercury as an air toxic; therefore, it was still subject to the MACT standards as laid out in section 112 of the Clean Air Act. Currently the US government is likely to reinstate MACT standards for mercury, but until then, mercury control will be left to the states, many of whom adopted stricter standards than the federal ones [44].

Mercury is a highly toxic substance, and its toxic effects have affected humans for millennia. Mercury produced in one corner of the world can affect another, which can make mercury contamination a political issue. The case of the change in US regulations highlights issues of scale and social justice: The cap and trade system may reduce the levels of mercury nationally, but it may do so at the expense of some local populations. That raises questions of whom regulations benefit (such as more well-off people at the expense of poorer who may live near power generation). It also illustrates the importance of scale when thinking about pollution and regulation: Nationally the levels of mercury might look acceptable, but when the resolution is increased and examined locally some places may have unacceptably high levels. This ties back to the issue mentioned above about how an air toxic is measured, when and where; different levels of measurement can have different results, and the action taken based on those measurements can affect health and outcomes.

## Human Health

Many of the health effects of different compounds that are considered air toxics are well established, and can be quite dramatic. In the early 1950s, residents of Minamata Bay, Japan began suffering from neurological illness. Adults began exhibiting symptoms such as blurred vision and difficulty in hearing, reduced ability to smell and taste, clumsiness, difficulty in walking, and mood changes. Children and newborns began suffering from impairments in chewing, swallowing, speech, and

movement. All of these individuals were suffering from brain damage due to acute poisoning with methyl mercury. Contamination came from a local acetaldehyde plant (acetaldehyde is used in the making of plastics). The residents became ill from eating fish and shellfish that had stored mercury in their bodies through biomagnification. Estimates put the number affected since the 1950s at 200,000 and these acute cases of illness have been unequivocally linked to mercury [45].

The results of air toxics in lower doses (the type one might encounter living next to a dry cleaner, or living in a heavy traffic prone area) is much more difficult to understand and to illustrate. Knowing that chemicals like benzene and PERC have carcinogenic effects in laboratory animals and during heavy occupational exposures is not the same as claiming that the average residents of a city are at greater risk for cancer because they live in a high traffic area. The relationship between toxics and illness is not always clear-cut. Benzene is one of the best understood of the air toxic carcinogens, and the effects have been well explored in animal research as well as in humans who are exposed to benzene in occupational concentrations. Less is known, however, about the relationship between ambient concentrations and cancer, especially in children. Whitworth et al., for example, used EPA models of benzene levels and census track to show that census tracks that have the highest levels of benzene also have higher incidence of childhood leukemia [20]. On the other hand, Reynolds et al. found that traffic density (which is a predictor of benzene and butadiene levels) did not correlate with childhood leukemia rates [46]. Thus even though the health effects of benzene are well established, the effects of environmental concentrations on populations are more difficult to pinpoint. With respect to mercury, Lewandowski et al. note that correlations have been found between mercury levels and autism in 2002, for school districts in Texas, but that these correlations were not found for 2005 and 2006, calling into question the original correlations [47]. Understanding the health effects is an important issue since policy and regulation decisions are made, at least in part (remember the EPA can create stricter regulations based on health effects) upon the impacts on human health. Understanding how these effects unfold in real-world situations is crucial to assessing local impacts of

pollutants, the benefits of regulation, and the need for new cleaner technologies.

The health effects of air toxics cannot be completely decoupled from general problems of air pollution. Often, sources of pollution, such as power plants and automobiles, release more than one harmful substance. In addition, multiple sources will cluster in certain areas, such as cities (many cities have power plants, drycleaners, and cars!). In areas that are developing quickly and do not have strict environmental controls, such as parts of Africa, individuals may be exposed to mercury, lead, pesticides, contaminated water, and various air toxics simultaneously [48]. Air toxics and other forms of pollution are also part of larger issues of health and illness: Factors such as several environmental pollutants, nutrition, levels of stress, and poverty interact and can be considered to affect health synergistically, with each problem causing the others to become more dangerous and vice versa [49, 50]. The impacts of toxics are complex and intertwined with the fates of other pollutants and linked to other aspects of human health and well-being. While it is important to understand how a particular toxin reacts, and how it may affect health, it is important not to lose sight of the larger system and the interactions that play a role in determining who and what is ultimately impacted.

## Future Directions

The regulation of air toxic chemicals is an evolving science. Not all aspects of transport and fate are completely understood (mercury is a good example of this). The activities that produce air toxics are part of the activities of daily life, and are difficult to terminate. People all around the world still need to get from place to place, and power must be generated. These needs have to be balanced with the impact from the consequences of these actions. The immediate strategy is to make current technologies (coal burning, gasoline automobiles, dry cleaning establishments) as clean as they can be. This presents challenges since the way in which toxics behave in the environment is not always completely understood (as is the case with mercury), regulations have costs to consumers and producers, and monitoring for toxics is still an underdeveloped science. In some cases, such as the push for cleaner

burning fuels, attempting to solve one problem (higher efficiency) leads to another (higher levels of volatile aromatics such as benzene). Some toxic-generating activities may be supplanted with newer technologies in future. This presents challenges as well. Changing from gasoline to electric powered cars will necessitate a change in the infrastructure that supports the automobile, and will create the need for collaboration between very different industries. Alternative sources of power, such as nuclear, have their own drawbacks in risks and wastes generated. Even alternatives such as wind have opponents who are concerned about the space needed to house wind generation and the impact on the aesthetics and enjoyment of land [51]. There are political issues at play as well. Much of the reason for the increase in coal burning is that it is a relatively inexpensive means of power generation for countries that are expanding rapidly. Many in those countries argue that they cannot afford to implement stringent environmental control technologies at the expense of the development that they need. The picture becomes more complex with toxics such as mercury, which can travel half way across the globe, so that the energy choices made by one country may affect another.

The way forward is likely a mixture of better monitoring and regulation of current technologies, as well as implementation of newer technologies that have less environmental impact. Many of these will likely exist in tandem (as is seen today with gasoline cars and gasoline electric “hybrid” cars). In order to create a way forward that is sustainable, solutions must be found that can benefit most people. The concerns of developed countries, developing countries, industry, and citizens must be considered, with the overall health and well-being of humans everywhere being the foremost concern. Cooperation between corporations and governments, as well as long-term thinking will be key to meeting these challenges.

## Bibliography

### Primary Literature

1. US EPA (2000) Taking toxics out of the air. Brochure of the US EPA Office of Air Quality Planning and Standards, Research Triangle Park, NC. Accessed 11/22/2010, Available from: <http://www.epa.gov/air/oaqps/takingtoxics/p1.html>

2. EPA (2007) The Plain English Guide to the Clean Air Act. 2008, US EPA Office of Air Quality Planning and Standards, Research Triangle Park, NC. Publication No. EPA-456/K-07-001
3. Keys D (2003) How Rome polluted the world. *Geogr Campion Interact Publishing* 75(12):45–48
4. Bell ML, Davis DL (2001) Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environ Health Perspect Suppl* 109:389
5. BeruBe K et al (2005) London smogs: why did they kill? *Proc R Microsc Soc* 40(3):171–183
6. Parris TM (2006) Smog Season. *Environment* 48(4):3–3
7. Aronson SM (2009) The ignoble fate of the peppered moth. In: *Medicine & Health Rhode Island*, pp 291–291
8. Corman R (1978) Air pollution primer. The American Lung Association, New York
9. US EPA (1994) Environmental fact sheet: air toxics from motor vehicles. In: *Environmental Protection Agency* (ed) Ann Arbor.
10. Stern RE (2003) Hong Kong Haze. *Asian Surv* 43(5):780–800
11. Fang G-C, Wu Y-S, Chang T-H (2009) Comparison of atmospheric mercury (Hg) among Korea, Japan, China and Taiwan during 2000–2008. *J Hazard Mater* 162(2/3):607–615
12. Luecken DJ, Cimorelli AJ (2008) Codependencies of Reactive Air Toxic and Criteria Pollutants on Emission Reductions. *J Air Waste Manage Assoc* 1995 58(5):693–701
13. McSwane D (2009) Food safety, in environmental health: from global to local. In: *Frumppkin H* (ed). Jossey-Bass, San Francisco
14. Logue JM et al (2009) High time-resolved measurements of organic air toxics in different source regimes. *Atmos Environ* 43(39):6205–6217
15. Raymer JH et al (2009) Microenvironmental characteristics important for personal exposures to aldehydes in Sacramento, CA, and Milwaukee, WI. *Atmos Environ* 43(25):3910–3917
16. Oñzkaynak H et al (2008) Modeling population exposures to outdoor sources of hazardous air pollutants. *J Expo Sci Environ Epidemiol* 18(1):45–58
17. Harley RA et al (2006) Effects of Reformulated Gasoline and Motor Vehicle Fleet Turnover on Emissions and Ambient Concentrations of Benzene. *Environ Sci Technol* 40(16):5084–5088
18. McCarthy MC et al (2009) Characterization of the Chronic Risk and Hazard of Hazardous Air Pollutants in the United States Using Ambient Monitoring Data. *Environ Health Perspect* 117(5):790–796
19. Raun LH, Marks EM, Ensor KB (2009) Detecting improvement in ambient air toxics: An application to ambient benzene measurements in Houston, Texas. *Atmos Environ* 43(20):3259–3266
20. Whitworth KW, Symanski E, Coker AL (2008) Childhood Lymphohematopoietic Cancer Incidence and Hazardous Air Pollutants in Southeast Texas, 1995–2004. *Environ Health Perspect* 116(11):1576–1580
21. US EPA (2009) Toxics Release Inventory (TRI) Program Fact Sheet. 2009 9/20/2009 [cited 2010 1/19]; Available from: [http://www.epa.gov/triinter/triprogram/tri\\_program\\_fact\\_sheet.htm](http://www.epa.gov/triinter/triprogram/tri_program_fact_sheet.htm)
22. US EPA (1996) The Aspen model. Technology Transfer Network 1996 National-Scale Air Toxics Assessment 2009 6/6/2007 [cited 2010 1/19]; Available from: <http://www.epa.gov/ttn/uatw/nata/aspen.html>
23. Venkatram A et al (2009) Modeling the impacts of traffic emissions on air toxics concentrations near roadways. *Atmos Environ* 43(20):3191–3199
24. Morrison B, Heath B (2009) States' tests of air questioned. *USA Today*
25. Heath B, Morrison B (2009) EPA study: 2.2 M live in areas where air poses cancer risk. *USA Today*
26. ATSDR (1997) American Toxic Substances and Disease Registry Toxfaqs: Tetrachloroethylene. US DHHS PHS ATSDR CAS#127-18-4
27. Graham LA, Belisle SL, Baas C-L (2008) Emissions from light duty gasoline vehicles operating on low blend ethanol gasoline and E85. *Atmos Environ* 42(19):4498–4516
28. US EPA (2006) Master list of compounds emitted by mobile sources. Environmental Protection Agency (ed) Washington, DC
29. Brown JL (2009) The Market Effects of Going Green: Evidence from California's Wholesale Reformulated Gasoline Market. *Energy J* 30(3):115–127
30. Tollefson J (2008) Car industry: Charging up the future. *Nature* 456(7221):436–440
31. EPA (2009) Fact sheet-final amendments to air standards for perchloroethylene dry cleaners. EPA, Editor
32. US EPA (2006) Proposed amendments to air toxics standards for perchloroethylene dry cleaners. Environmental Protection Agency, Editor
33. US EPA (1993) Air pollutants, hazardous; national emission standards: perchlorethylene emissions from dry cleaning facilities. Edited by Environmental Protection Agency, Research Triangle Park, NC, p 49354
34. Department of Health and Human Services (1999) Agency for toxic substances and disease registry, *Mercury*, Department of Health and Human Services, Atlanta
35. Sun L et al (2006) A 2000-year record of mercury and ancient civilizations in seal hairs from King George Island, West Antarctica. *Sci Total Environ* 368(1):236–247
36. Li P et al (2009) Mercury pollution in Asia: a review of the contaminated sites. *J Hazard Mater* 168(2–3):591–601
37. Tillett T (2009) The Price of Progress. *Environ Health Perspect* 117(6):A257–A257
38. Levin P (2007) From mad hatters to dental amalgams: heavy metals: toxicity and testing. *MLO Med Lab Obs* 39(12):20
39. Selin NE (2009) Global Biogeochemical Cycling of Mercury: A Review. *Annu Rev Environ Resour* 34:43–63
40. Streets DG, Zhang Q, Wu Y (2009) Projections of Global Mercury Emissions in 2050. *Environ Sci Technol* 43(8):2983–2988
41. Glodek A, Pacyna JM (2009) Mercury emission from coal-fired power plants in Poland. *Atmos Environ* 43(35):5668–5673
42. EPA (2002) National Risk Management Research Laboratory Research, Triangle Park, NC. Control of mercury emissions

from coal-fired electric utility boilers: interim report including errata dated 3-21-02

43. Pudasainee D, Kim JH, Seo YC (2009) Mercury emission trend influenced by stringent air pollutants regulation for coal-fired power plants in Korea. *Atmos Environ* 43(39): 6254–6259
44. Milford JB, Pieniak A (2009) After the Clean Air Mercury Rule: Prospects for Reducing Mercury Emissions from Coal-Fired Power Plants. *Environ Sci Technol* 43(8):2669–2673
45. Ekino S et al (2007) Minamata disease revisited: An update on the acute and chronic manifestations of methyl mercury poisoning. *J Neurol Sci* 262(1–2):131–144
46. Reynolds P et al (2002) Traffic patterns and childhood cancer incidence rates in California, United States. *Cancer Causes Control* 13(7):665–673
47. Lewandowski TA et al (2009) An Evaluation of Surrogate Chemical Exposure Measures and Autism Prevalence in Texas. *J Toxicol Env Health A* 72(24):1592–1603
48. Nweke OC, Sanders WH III (2009) Modern Environmental Health Hazards: A Public Health Issue of Increasing Significance in Africa. *Environ Health Perspect* 117(6):863–870
49. Singer M, Clair S (2003) Syndemics and Public Health: Reconceptualizing Disease in Bio-Social Context. *Med Anthropol Quart* 17(4):423–441
50. Goodman AH, Leatherman TL (eds) (1998) Building a new biocultural synthesis. University of Michigan Press, Ann Arbor, pp 3–42
51. Johansson M, Laike T (2007) Intention to respond to local wind turbines: the role of attitudes and visual perception. *Wind Energy* 10(5):435–451

## Books and Reviews

- Ganesan K et al (1996) *Air toxics: problems and solutions*. Gordon & Breach, Amsterdam, p 342
- Griffin RD (2007) *Principles of air quality management*, 2nd edn. Taylor and Francis, London, 334 pp
- Hanna SR et al (1982) *Handbook on atmospheric diffusion*. Office of Energy Research, US Department of Energy, 102 pp
- Harrop O (2002) *Air quality assessment and management: a practical guide*. Taylor & Francis, London, 504 pp
- Heinsohn RJ, Kabel RL (1999) *Sources and control of air pollution*. Pearson Education, Upper Saddle River, p 696
- Jacobson MZ (2000) *Atmospheric pollution: history, science and regulation*. Cambridge University Press, New York, p 399
- Spicer CW et al (2002) *Hazardous air pollutant handbook: measurements, properties, and fate in ambient air*. Lewis, Boca Raton, p 240
- Turner DB (1994) *Workbook of atmospheric dispersion estimates: an introduction to dispersion modeling*. Lewis, Boca Raton
- USEPA (2004) *Air toxics risk assessment reference Library*, edited by Office of Air Quality Planning and Standards, EPA, Research Triangle Park, NC
- Vallero D (2008) *Fundamentals of air pollution*, 4th edn. Academic Press, San Diego, 942 p

## Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL)

RUSSELL DE YOUNG

NASA Langley Research Center, Hampton, VA, USA

### Article Outline

Glossary  
Definition of the Subject and Its Importance  
Introduction  
Future Directions  
Bibliography

### Glossary

**Absorption** The process in which incident radiant energy is retained or absorbed by a gas.

**Absorption coefficient** The fraction of incident energy removed by absorption per unit length of travel of radiation through a gas.

**Active remote sensing** A remote sensing system that transmits its own energy source and measures atmospheric properties from the returned signal.

**Aerosol** A colloidal suspension of liquid or solid particles in air.

**Aerosol extinction** The reduction of optical energy passing through the atmosphere due to scattering and absorption of that energy.

**Atmospheric boundary layer** The bottom layer of the atmosphere that is in contact with the surface and is often turbulent with a capped stable layer of air or temperature inversion. The height is variable but is typically several kilometers during the day.

**Depolarization ratio** Defined as the ratio of the backscattered lidar power in the cross-polarization plane to the backscattered power in the polarization plane (transmitted laser beam) induced by the atmospheric constituents.

**DIAL** Differential absorption lidar is a lidar technique used to measure the atmospheric concentration of any gas as a function of altitude.

**Extinction coefficient** The fraction of incident radiant energy removed by extinction per unit of travel of radiation through the air.

**In situ** Methods for obtaining information about atmospheric properties through direct contact with the atmosphere as opposed to remote sensing.

**Lidar** Light Detection and Ranging is a technique for detecting and characterizing atmospheric properties using a pulsed laser.

**Passive remote sensing** A remote sensing technique that relies on the use of the Sun, Moon, or Earth surface as a radiation source to measure atmospheric gas species by absorption of that radiation as it passes through the atmosphere.

**Polarization** The property of light where the electric field vector is oriented in a single plane called plane polarization.

**Primary trace atmospheric gas or particles** Substances that are directly emitted into the atmosphere from the surface, vegetation, or natural or human activities such as fires, industrial processes, and car emissions.

**Relative humidity** The ratio of the vapor pressure of water to its saturation vapor pressure at the same temperature.

**Scattering** The process by which atmospheric gases are excited to radiate by an external source of light and the resultant light is usually detected in a direction not aligned with the light source.

**Scattering coefficient** The fraction of incident radiant energy removed by scattering per unit length of travel of radiation through the air.

**Stratosphere** The region of the atmosphere from the top of the troposphere to a height of 10–17 km (the base of the mesosphere).

**Troposphere** The region of the atmosphere from the Earth's surface to the tropopause at 10–20 km, depending on latitude and season, where most weather occurs.

## Definition of the Subject and Its Importance

The study of the atmosphere has expanded greatly in the past decade due to concern about global climate change and air quality health effects. The natural atmospheric chemistry is complex but with anthropogenic emissions released into the atmosphere, the resulting complexity makes modeling very difficult. Chemical reactions generally increase with temperature and thus a warming climate may change the weather and

climate in unpredictable ways. Also with increased regulations regarding air quality emissions, more attention is being directed toward atmospheric species measurements to assess the impact of specific emission regulations. As a result of these concerns, lidar has become a very valuable tool to directly measure the number density of specific atmospheric species as a function of altitude. This article will review the use of differential absorption lidar (DIAL) in current aircraft based missions and the potential for use of DIAL in orbiting spacecraft

A particular lidar technique called Differential Absorption Lidar (DIAL) has proven to be an especially powerful method to measure atmospheric species. It has been used both on the ground and in many aircraft to determine the concentration of gas species. There are presently no DIAL systems in space, but there are projections by NASA and other nations to place them in Earth orbit to measure critical gases over remote areas and for long periods of time. This article will discuss the DIAL technique, demonstrate its usefulness using airborne examples, and discuss the value of a space-based DIAL satellite system.

## Introduction

The need for accurate atmospheric gas concentration measurements as a function of altitude has expanded as a result of predicted atmospheric changes associated with global climate change and the impact of air quality on human health. Both local natural and anthropogenic emissions can be transported around the Earth and thus have health impacts far from the emission sources. Lidar has proven to be a powerful technique for measuring the concentrations of many gas species related to health and climate change concern.

Differential Absorption Lidar (DIAL) from aircraft was first accomplished for ozone at NASA Langley Research Center in 1980 [1] and for water vapor in 1982 [2]. Currently DIAL systems are being flown by NOAA and European countries, but presently there are no DIAL systems deployed in space.

The DIAL system has the ability to address major questions concerning the atmosphere. DIAL systems can be deployed in fixed ground locations, in either piloted or unpiloted aircraft or in satellite platforms. Each has distinct advantages for addressing



atmospheric research questions. Some of these questions are:

- How does urban pollution transport out of the boundary layer into the free troposphere and then to continental scales?
- How do atmospheric species chemically transform with transport age and reaction with other species?
- What vertical and horizontal resolution measurements are needed to substantially improve chemical weather models?
- What atmospheric species need to be measured simultaneously to understand atmospheric chemical transformation?
- What are the temporal and spatial water vapor scales needed to improve hurricane intensification and path forecasting?
- To what resolution do we need to measure atmospheric water vapor to make future weather predictions for every square kilometer and for 1 month in the future for the entire globe?
- Can climate change impacts be measured for chemical weather?

### Atmospheric Backscattering and the DIAL Equation

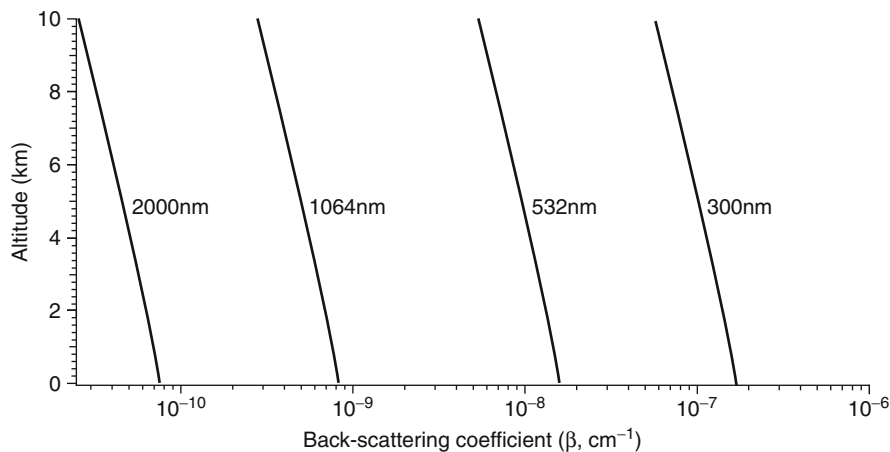
The ability to use laser remote sensing relies on the characteristic of electromagnetic radiation scattering from atmospheric aerosols and molecules. The atmosphere consists primarily of nitrogen and oxygen

molecules as well as highly variable aerosols. Scattering from molecules or Rayleigh scattering can be calculated as [3]

$$\beta = (3/8\pi)\beta_s(P/P_s)(T_s/T) \quad (1)$$

where  $\beta_s$  is the volume scattering coefficient ( $\text{km}^{-1}$ ) for the laser wavelength used,  $P_s$  is 1013.25 millibars,  $T_s$  is 288.15 K, and  $T$  and  $P$  are the standard pressures and temperatures at altitude. Equation 1 describes the backscatter coefficient that scatters some laser radiation back to the DIAL receiver system. Figure 1 shows the results of Eq. 1 for several common laser wavelengths. The wavelengths at 355, 532, and 1,064 nm are associated with the Nd:YAG laser system while 2,000 nm is commonly used for CO<sub>2</sub> DIAL measurements. Several important properties are noted from the figure. The backscatter coefficient decreases by three orders of magnitude from 355 to 2,000 nm making the use of DIAL more difficult in the infrared. The backscatter is much higher at lower altitudes due to the increased atmospheric density but this also has a disadvantage in that optical detectors are easily saturated by the near field atmospheric lidar return causing signal-induced noise. In contrast the backscatter at 10 km is much less, but this is an advantage for aircraft deployed systems in that the near field signal is substantially reduced, thus reducing signal-induced noise in optical detectors [4].

Scattering also occurs from atmospheric aerosols, but aerosols are highly variable in spatial and temporal



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 1**  
Backscattering coefficient vs. altitude for different laser wavelengths and northern latitude standard atmosphere

extent. Nevertheless, both aerosols and molecules contribute to the lidar backscatter signal.

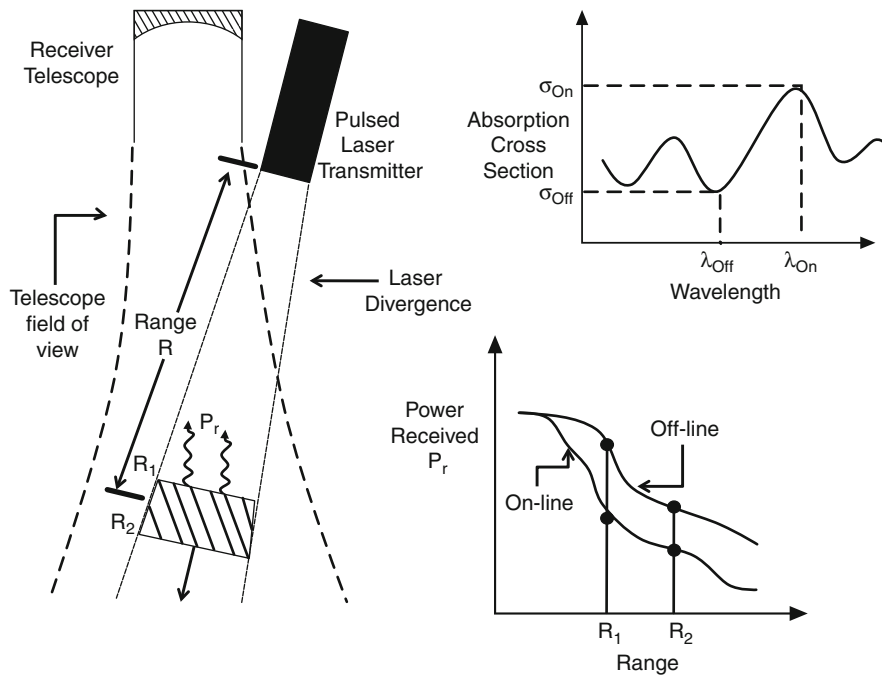
The DIAL equation, giving the atmospheric number density  $N$  of any gas species, is derived from the lidar equation which is discussed in detail in [5–7] and given as

$$\begin{aligned}
 N(R) = & 1/(2(R_2 - R_1)(\sigma_{\text{on}} - \sigma_{\text{off}})) \\
 & * \ln(P_{\text{on}}(R_1)P_{\text{off}}(R_2)/P_{\text{off}}(R_1)P_{\text{on}}(R_2)) \\
 & - 1/(2(R_2 - R_1)(\sigma_{\text{on}} - \sigma_{\text{off}})) \\
 & * \ln(\beta_{\text{on}}(R_1)\beta_{\text{off}}(R_2)/\beta_{\text{off}}(R_1)\beta_{\text{on}}(R_2)) \\
 & - 1/(\sigma_{\text{on}} - \sigma_{\text{off}}) * (\alpha_{\text{on}} - \alpha_{\text{off}})
 \end{aligned} \quad (2)$$

where the third term is a correction for scattering and absorption of interfering gas species as well as the finite difference (due to the on- and off-line wavelength difference) in the average extinction coefficient,  $\alpha$ . If care is taken on the choice of wavelengths to avoid interfering gases and the difference between on- and off-line wavelengths is small, then this term is negligible. The second term accounts for an inhomogeneous atmosphere where the backscatter coefficient,  $\beta$ , is

different at ranges  $R_1$  and  $R_2$ . Assuming a homogeneous atmosphere and a sufficiently small on- and off-line wavelength difference, then this term can be made negligible. The first term is the traditional DIAL equation. Here  $R_2 - R_1$  is the range cell defined by the digitization rate of the receiver digitizer. The differential absorption cross section,  $\sigma_{\text{on}} - \sigma_{\text{off}}$ , is the difference between the on- and off-line cross sections at the on- and off-line wavelengths. The “ln” term is the ratio of the on- and off-line powers received at either  $R_1$  or  $R_2$  ranges. Kuang [8] gives an example of a typical ozone DIAL analysis for a ground-based system with an aerosol correction procedure.

Figure 2 outlines the characteristics of the DIAL technique. A pulsed laser emits pulses at either the on- or off-line wavelength with a narrow divergence. The range of interest,  $R$ , is determined by  $ct/2$  where  $c$  is the light velocity and  $t$  is the round trip time. At some point in the atmosphere, the laser divergence intersects the receiver field of view and due to atmospheric scattering, receives the power  $P(r)$ . The range cell is usually set by the digitization rate, for example a rate of 5-MHz results in a 30-m range cell ( $R_2 - R_1$ ). The online



Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 2 Characteristics of the DIAL technique

wavelength corresponds to a well-known absorption feature of the particular gas species to be measured. Since the linewidth of the laser has some finite width, care is taken to determine the convolution of the laser linewidth with the absorption linewidth. The off-line is chosen near the online wavelength but in a region where there is no (little) absorption by the gas species of interest. Generally, the on- and off-line pulses are alternating at a repetition rate as high as possible to allow noise reduction by averaging without sacrificing horizontal resolution. The on- and off-line powers at each range cell  $R_2$ – $R_1$  are recorded at the receiver computer for postprocessing into curtain profiles of the particular gas species.

Several examples will now be described that illustrate the power of the DIAL technique to profile the atmosphere for different species from both aircraft and space platforms.

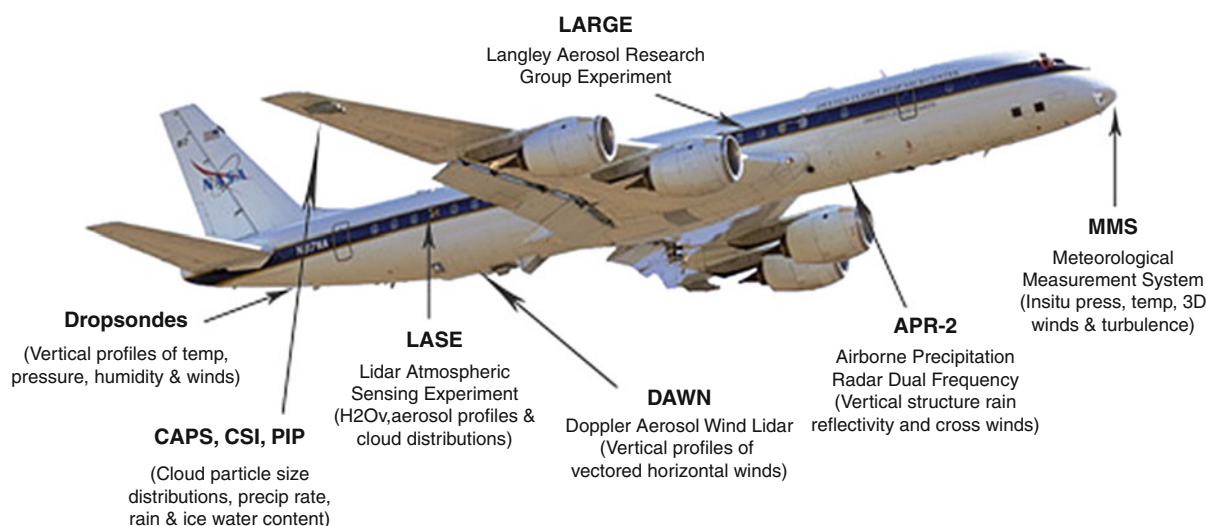
### Airborne Ozone DIAL

NASA has a variety of aircraft that are regularly involved in atmospheric research. The most capable aircraft is the NASA DC-8 aircraft shown in Fig. 3, which has flown in research mission all over the world for the past 20 years. The DC-8 carries an ozone or water vapor DIAL system along with

a dozen other instruments that probe the atmosphere resulting in a synergistic measurements platform.

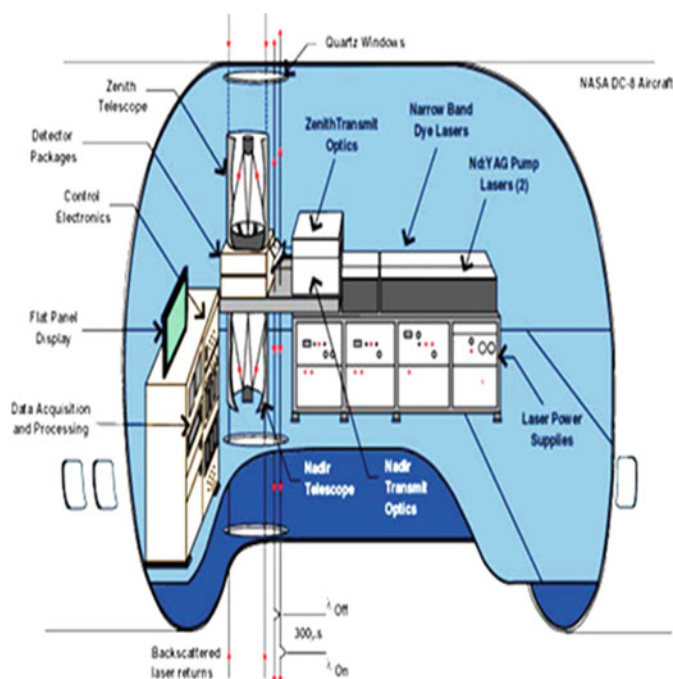
Figure 4 shows a schematic of the NASA Langley ozone DIAL system on the DC-8 aircraft. There are two Nd:YAG lasers: one for the online (290 nm) and the other for the off-line (300 nm) wavelength. Each Nd:YAG laser pumps a corresponding dye laser which in turn is doubled into the ultraviolet spectral region at 300 or 290 nm with a pulse repetition rate of 40 Hz each and an on- and off-line separation of 350  $\mu$ s. The on- and off-line laser pulse energy is split in two: One is transmitted in the zenith and the other in the nadir. As the aircraft flies, ozone profiles from the ground to the aircraft and from the aircraft to the stratosphere are simultaneously recorded. These two profiles can then be spliced together with the aircraft in situ ozone data, resulting in a complete ozone atmospheric profile. The ozone is usually presented in terms of mixing ratio or parts per billion by volume (ppbv). The vertical return signal is averaged to produce a resolution of 300 m and the horizontal resolution is 12 km.

The DIAL technique also has the advantage of retrieving data on clouds and aerosols. The residual dye laser energy at 599 nm (zenith) and at 583 nm (nadir) is vertically polarized and along with the residual 1,064-nm energy is transmitted with the UV ozone wavelengths. The visible wavelengths are received and



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 3**

The NASA DC-8 aircraft with several measurement systems listed including the water vapor DIAL system LASE



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 4**  
Schematic and photo of the NASA Langley ozone DIAL system on the DC-8 aircraft

sent through a cross polarizer to determine the percentage of backscattered depolarization. The resolution of the aerosol channel is 60 m in the vertical and 400 m in the horizontal.

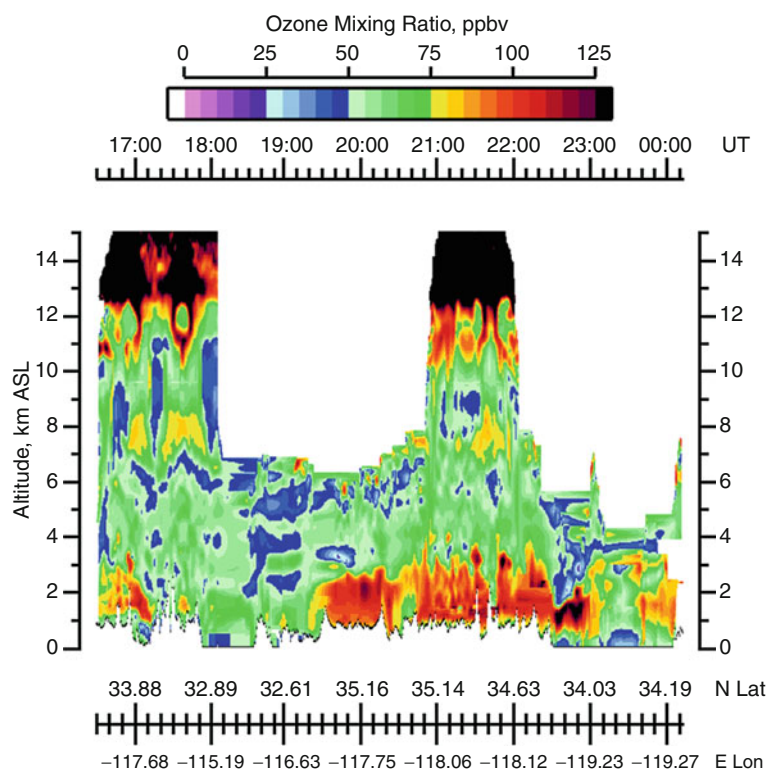
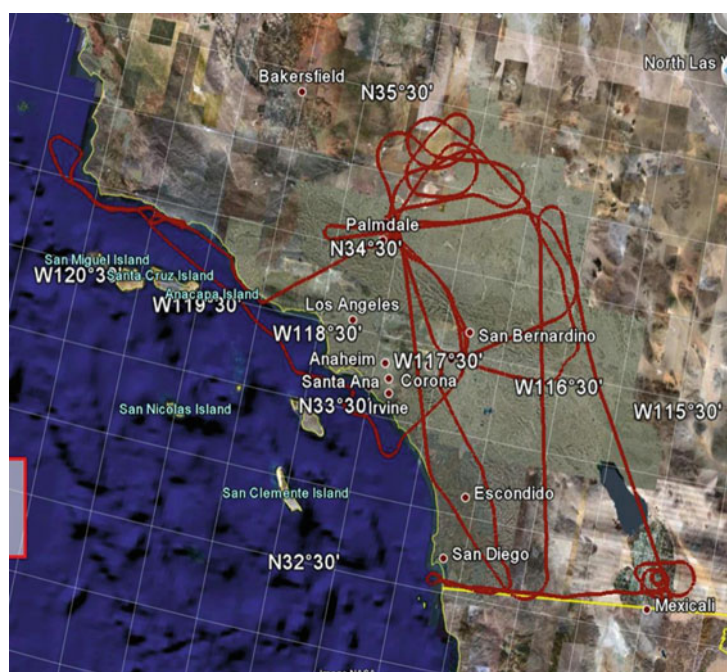
An example is given to demonstrate the science derived and the capability of the DIAL technique for measuring atmospheric ozone. This example is from the NASA Arctic Research of the Composition of the Troposphere from Aircraft and Satellites (ARCTAS) mission [9] and presents a complete DIAL data set for flight 15 on June 24, 2008.

Figure 5 shows the 8-h DC-8 flight track around Los Angeles, CA. The corresponding ozone atmospheric profile is shown for the flight track. The ozone concentration, as a function of altitude, is colored and corresponds to an ozone mixing ratio of 0–125 ppbv. This profile is interesting as it shows both stratospheric and boundary layer ozone. In the stratosphere the ozone is typically greater than 2,500 ppbv, so the figure shows the lower edge of the stratospheric ozone layer that protects life from harmful UV exposure from the Sun. The dynamics of the stratospheric/tropospheric boundary cause ozone filaments to bleed into the

troposphere at 12 km as shown in the figure. The ground profile is shown in kilometers above sea level and in the mountains north of Los Angeles high ozone concentrations are observed. At the flight time, there were numerous fires in the mountain region generating both aerosols and ozone. Ozone concentrations above 100 ppbv are measured and seen to transport out into the Pacific Ocean where the ozone decreases as the flight track moves northwest. While this ozone was generated in remote areas, the ozone generated in urban areas as a result of anthropogenic emissions can have severe health effects especially above background levels of 40 ppbv (U.S. EPA primary ozone standard is 75 ppb) and DIAL has the ability to measure the atmospheric distribution of this ozone. Likewise, stratospheric ozone has been measured especially with regard to stratospheric depletion in Antarctica due to chlorofluorocarbon emissions.

The residual laser energy at 1,064 and 591 nm (average of the two dye laser wavelengths) is transmitted along with the DIAL UV wavelengths to probe the aerosol distribution above and below the aircraft. If the aerosol size parameter  $2\pi r/\lambda < 1$ , where  $r$  is the aerosol radius





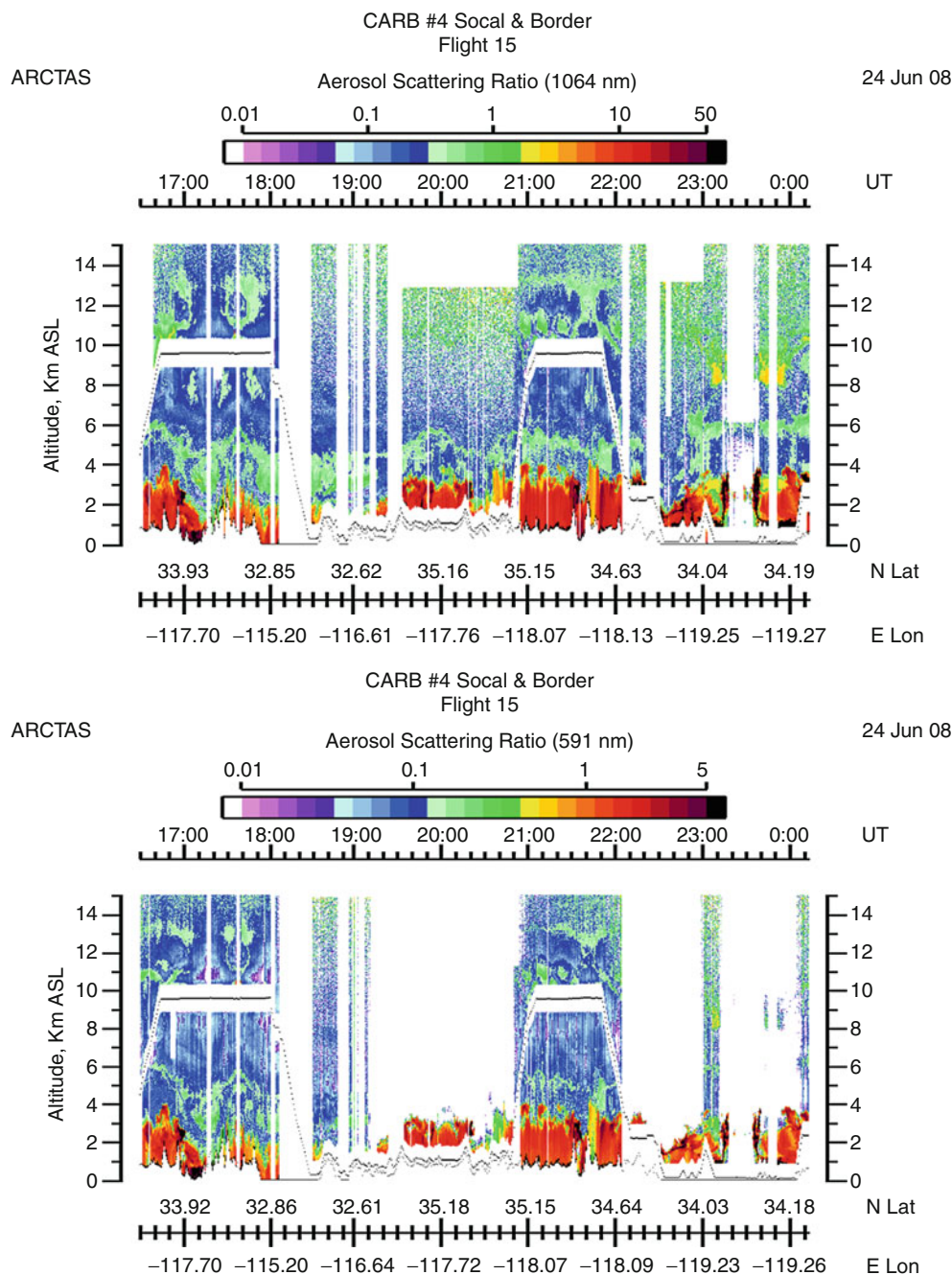
**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 5**

The aircraft flight plan for ARCTAS campaign on June 24, 2008 and the corresponding DIAL ozone atmospheric profile in parts per billion by volume



and  $\lambda$  the lidar wavelength, then large backscattering occurs. The 1,064- and 591-nm wavelengths probe different aerosol size distributions. Figure 6 shows the aerosol backscattering ratio at both 1,064- and 591-nm

wavelengths as a function of altitude. This ratio is the backscatter at either wavelength to the molecular atmospheric backscatter at the pressure altitude. If the ratio is zero, then there is no aerosol scattering, only



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 6**

The aerosol scattering ratio for 1,064 and 591 nm (average of the dye laser wavelengths) as a function of altitude. The ground profile above sea level is noted as well as the aircraft flight track

molecular scattering. Most of the aerosol backscattering occurs near the ground which is the atmospheric source of aerosols. In the figure, high aerosol scattering occurs near the fire sites in the mountains but aerosols are also transported out over the ocean (at times past 22.50 UT) where they are hydrated by moisture and thus grow in size producing greater backscatter. It is interesting to note that there are no clouds detected in the profile, yet nowhere in the atmosphere is there pure molecular scattering. This has implications for climate change as aerosols can either reflect or absorb incident energy, affecting the energy balance of the global atmosphere.

The aerosol wavelength dependence (color ratio) is shown in Fig. 7. The color ratio is the ratio of 1,064 to 591-nm backscattering and gives an indication of relative aerosol size with larger ratios indicating larger aerosols. As seen during the flight time of 21:00–22:00, finer particles are emitted by these fires but as they rise they grow in size, probably due to hydration. The aerosols also tend to grow in size as they transport over the ocean after time 22:30.

Figure 7 also shows the aerosol depolarization at 591 nm derived from the aerosol scattering ratio. The laser is linearly polarized as it is transmitted into the atmosphere and two receiver channels are used to detect the atmospheric return, one capturing the linear polarized return and the other the orthogonal or depolarized return signal. The ratio of the depolarized signal to the transmitted linear polarized signal is the aerosol depolarization. Spherical aerosols backscatter light with little polarization change, but nonspherical aerosols generally depolarize the backscatter signal depending on their shape [10]. Thus, the depolarization gives some information on the structure and shape of the atmospheric aerosols and can distinguish between liquid and solid phase aerosols. In the figure, the depolarization measured was generally between 3% and 9%. The fires at 17:00 UT appear to contain less dust (less depolarization) than fires at 21:00–22:00 UT, where the depolarization is higher, indicating the fires there lifted more dust into the fire plume.

This example of the DIAL technique for measuring ozone demonstrates the utility of DIAL. Here both the stratospheric and tropospheric ozone concentration can be determined simultaneously. Further the residual laser energy of the dye lasers is used to probe the

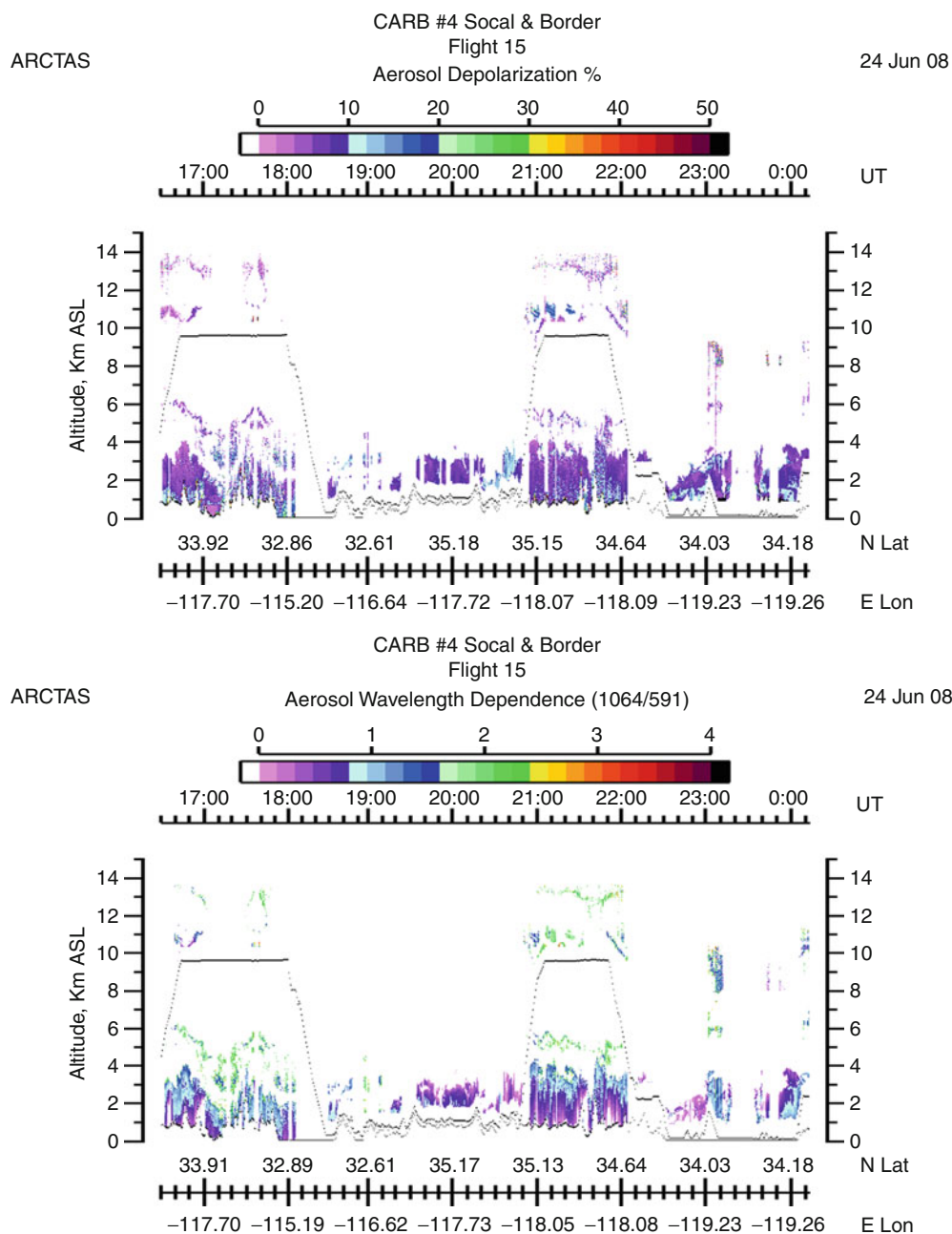
aerosols at 591 and 1,064 nm, showing the importance of aerosols in the boundary layer as well as aerosol transport in the troposphere. This example shows the importance of measuring both ozone and aerosols simultaneously.

The ozone absorption cross section is broad in the ultraviolet spectral region (250–320 nm), allowing many different wavelength on- and off-line pairs to be used. Absorption cross section for ozone and a number of atmospheric gases of interest for atmospheric chemistry are shown by Schouepnikoff [11]. For the stratosphere where ozone density is large, wavelength pairs near 320 are used in order to generate a significant return signal. Whereas in the troposphere where the ozone mixing ratio is lower, higher absorption is needed, thus wavelength pairs from 280 to 300 are used.

### Airborne Water Vapor DIAL

Atmospheric water vapor is the most radiatively active gas in the atmosphere and as a result affects weather patterns, precipitation, clouds, and hurricanes. Water vapor and ozone affect the production of OH radicals which help determine tropospheric chemistry. At high relative humidities, aerosols grow hygroscopically and in turn reduce visibility. Water vapor has very high temporal and spatial variability in the atmosphere and radiosonde or space-based satellite passive instruments cannot fully characterize the global variability. Airborne and eventually space-based water vapor DIAL systems could characterize the water vapor distribution for climate change and real-time weather forecasting by providing high vertical and horizontal profiles of water vapor distributions for input into climate models. Like ozone DIAL, a water vapor DIAL system can also measure clouds and aerosols, which significantly improves our understanding of water vapor distribution dynamics. Clouds and aerosols are currently the largest source of uncertainty in global climate models and both play a major role in atmospheric chemistry.

Figure 8 shows a schematic and photo of the Laser Atmospheric Sensing Experiment (LASE) water vapor DIAL system on the NASA DC-8 aircraft. The LASE system uses a Ti:sapphire-based laser transmitter [12]. This laser is a 5-Hz double-pulsed (300- $\mu$ s separation)

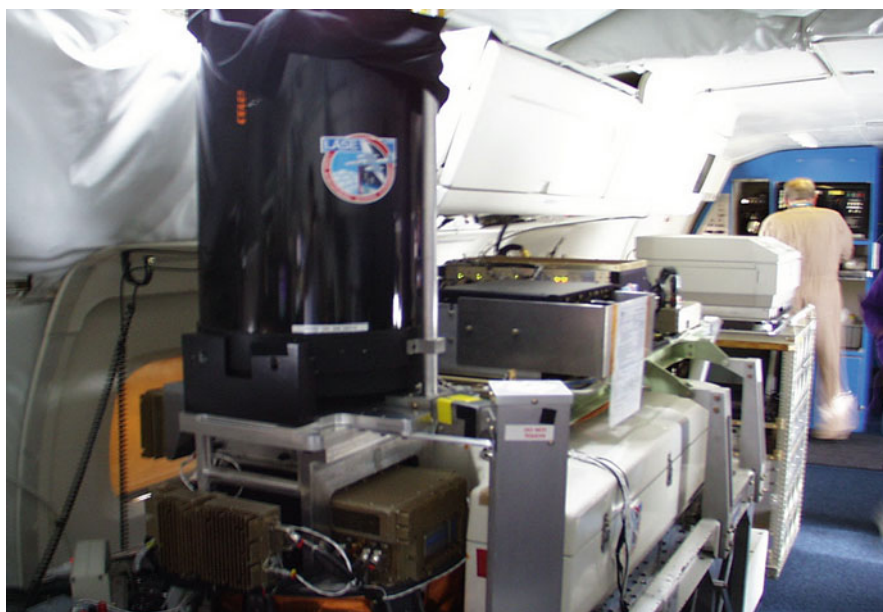
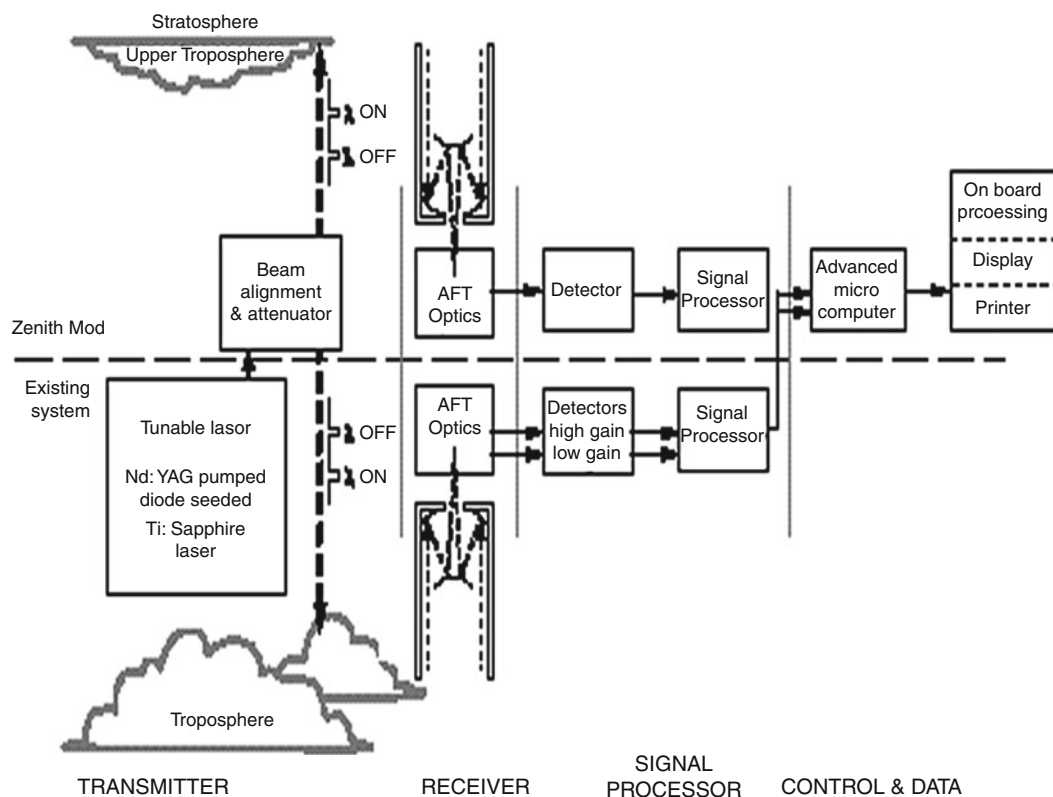


**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 7**

The percent aerosol depolarization and the ratio of 1,064–591 nm backscatter is shown as a function of altitude

Ti:sapphire laser that operates on the 817.2213-nm (vac) absorption line of water vapor and is pumped by a frequency-doubled flashlamp-pumped Nd:YAG laser. The laser wavelength is varied along the water vapor absorption line to create an on- and off-line

DIAL wavelength pair. LASE can operate over two or three water vapor concentration regions to cover a large altitude region (dynamic range) in the troposphere. This unique method of operation permits rapid and more flexible absorption cross-section selection



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 8**  
The Laser Atmospheric Sensing Experiment (LASE) water vapor DIAL system schematic and photo of instrument on the NASA DC-8 aircraft

capability for water vapor measurements over the entire troposphere in a single pass. The off-line wavelength is used to measure the aerosol scattering ratio and is selected for minimum absorption by water vapor. The wavelength of the laser is controlled by diode laser injection seeding that is frequency locked to the water vapor line using a water vapor absorption cell. The laser linewidth is less than 0.25 pm and the line stability is approximately 0.35 pm, much less than the pressure broadened atmospheric water vapor linewidth. The pulse energy ranges from 75 to 100 mJ and approximately 60% of the energy is transmitted in the nadir direction and 40% in the zenith direction. The beam divergence is typically 1 mrad and the telescope field of view is 1.5 mrad. The detector system consists of two silicon avalanche photodiodes (Si:APD) and three digitizers to cover a large signal dynamic range ( $10^6$ ). The data system on the DC-8 will enable real-time and postflight analyses onboard the aircraft.

Atmospheric water vapor profiles are taken during night or day with a vertical resolution of 330 m and horizontal resolution of from 14 to 42 km. The aerosol resolution is 30 m in the vertical and 200 m in the horizontal. The accuracy of the water vapor measurement is 6% or 0.01 g/kg whichever is greater. The atmospheric in situ water vapor aircraft data can be spliced into the LASE DIAL data to produce a complete water vapor profile throughout the troposphere.

Figure 9 is a profile example from the Genesis and Rapid Intensification Processes (GRIP) DC-8 aircraft campaign from August 15 to September 30, 2010 [13]. This flight segment took place east of Denver, CO over a relatively flat agriculture area. In the figure, the boundary layer is shown below 2 km containing significant amounts of water vapor. Above the boundary layer the air mass is quite dry. Back trajectories have determined that this air mass came from the California fires occurring at that time, thus there is enhanced backscattering from smoke particles at 5 km. The aerosol scattering ratio is increased at the top of the boundary layer due to absorption of water by the aerosols, increasing their size, thus producing this thin layer of enhanced backscattering.

Figure 10 outlines the analysis process by which the DIAL raw data is transformed into useful data products [14]. The atmospheric return signals are combined with ground elevation data and DC-8 navigation and

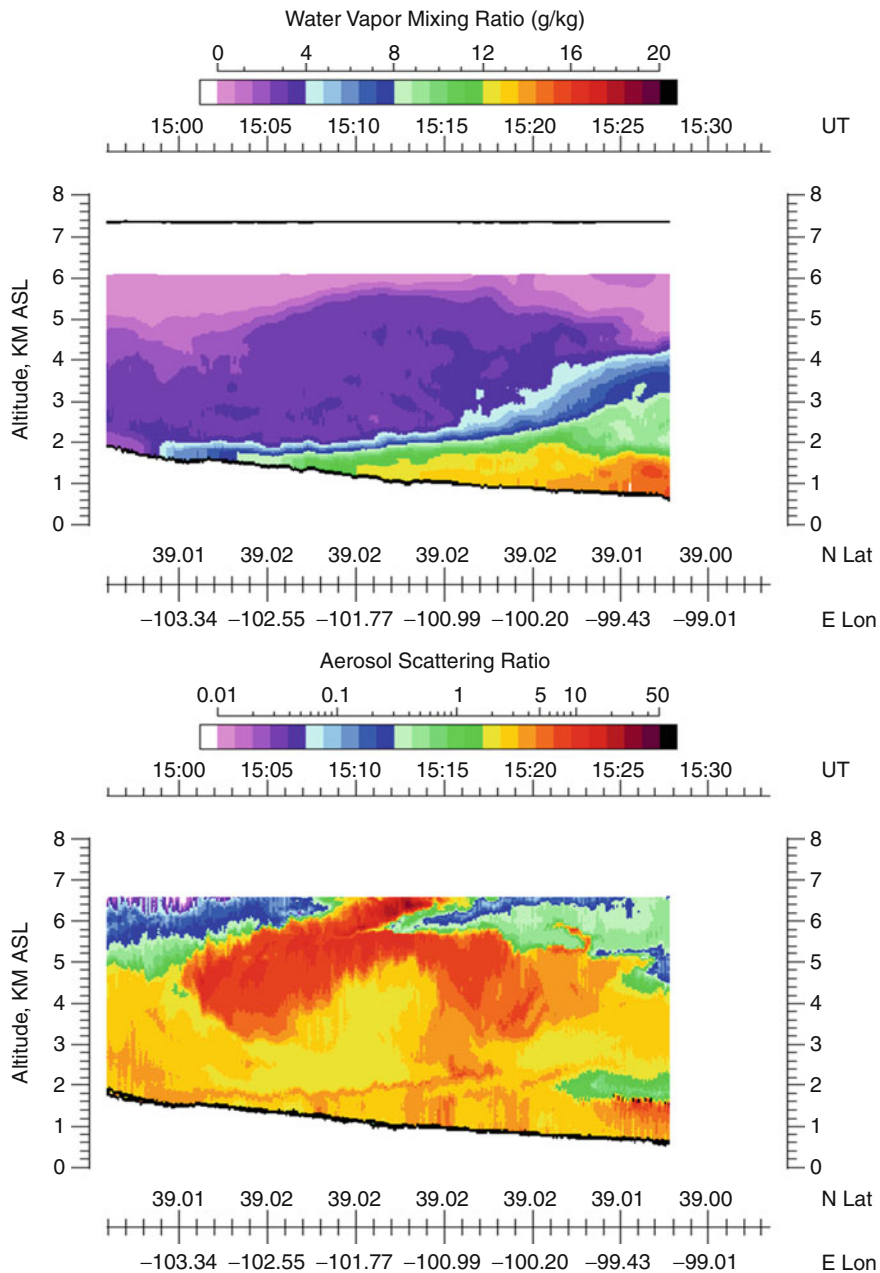
in situ water vapor data to create a composite water vapor return. There are three wavelength on- and off-line pairs used for different atmospheric water absorption conditions as well as three amplifier gains to increase the measurement dynamic range. This data is used to create an initial DIAL calculation for water vapor and also relative aerosol backscatter. Temperature, pressure, and relative humidity data from radiosondes on the DC-8 are used to calculate the effect of temperature and pressure on water vapor line broadening as well as aerosol scattering. The nadir and zenith profiles are combined with in situ water vapor data to produce a complete water vapor profile of the troposphere. The aerosol scattering analysis is done in an iterative process to create the aerosol scattering ratio with water correction.

### Space-Based DIAL

Currently there are no space deployed Earth orbiting DIAL systems. There are simpler backscatter Earth orbiting lidar systems to measure clouds and aerosols. The first of these was the Laser In-space Technology Experiment (LITE) which flew for 10 days on the Space Shuttle in 1994 [15]. This backscatter lidar used a flashlamp-pumped Nd:YAG laser with nonlinear conversion into 532 and 355 nm laser emission. Three beams were transmitted (1,064, 532, and 355 nm) and provided the first highly detailed global view of the vertical structure of clouds and aerosols from the surface through the middle stratosphere. The complexity of deploying a lidar in space is a daunting challenge and LITE performed extremely well demonstrating the utility of backscatter lidars in space even though its deployment was very short.

Another space-based lidar was the Ice Cloud and land Evaluation Satellite (ICESat) which was launched in 2003 and decommissioned in 2010 [16, 17]. The ICESat mission was designed to provide elevation data needed to determine ice sheet mass balance as well as cloud property information, especially for polar stratospheric clouds and aerosols. It provided topography and vegetation data around the globe, in addition to the polar-specific coverage over the Greenland and Antarctic ice sheets. It also provided planetary boundary layer heights which is an important parameter linking the surface to atmosphere dynamics, especially

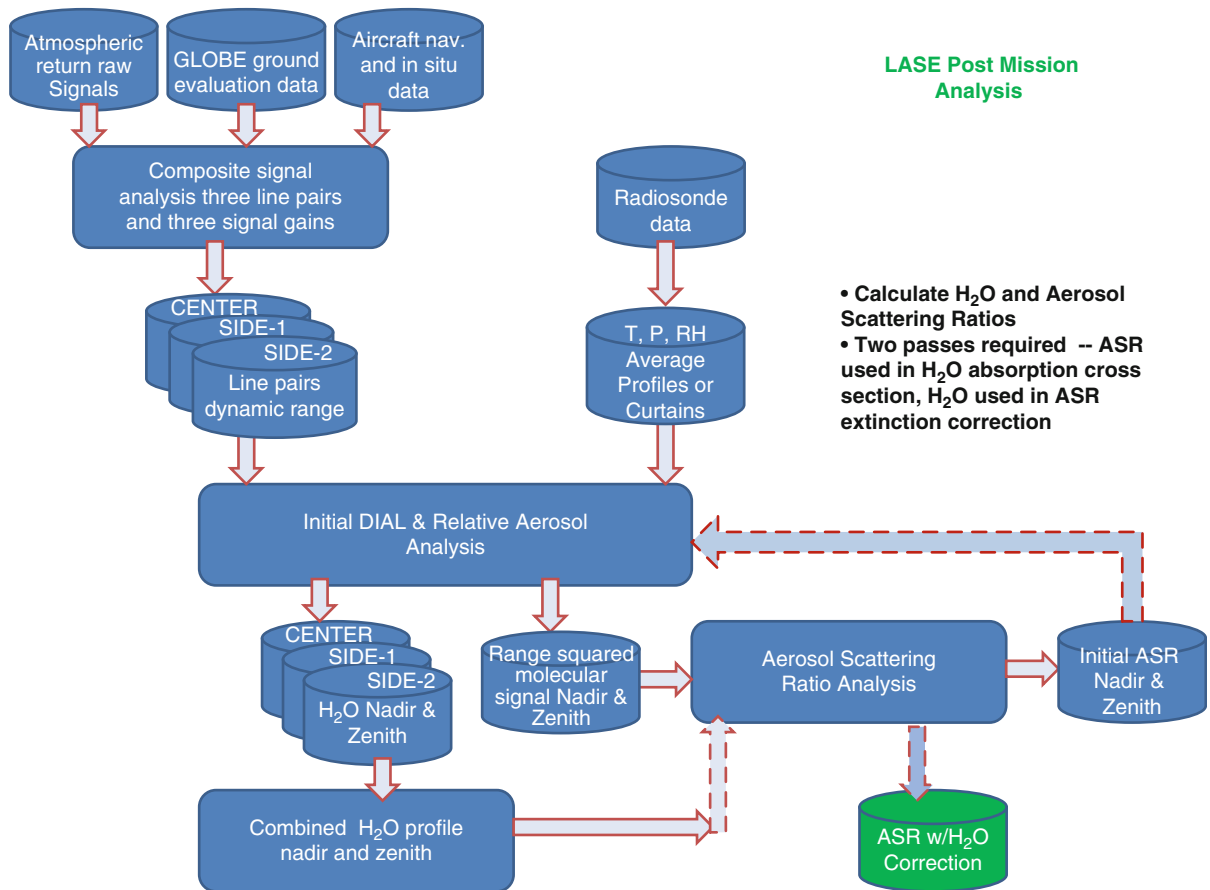




**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 9**  
 Water vapor mixing ratio and the corresponding aerosol scattering ratio during the GRIP campaign [12]

over the world's oceans. The satellite combined a precision surface lidar with a sensitive dual-wavelength cloud and aerosol lidar. The lasers emitted infrared and visible laser pulses at 1,064 (75 mJ) and 532 (35 mJ) nm wavelengths using three identical diode-pumped Nd:YAG lasers (6-ns TEM00 spatial mode). The repetition

rate was 40 Hz. There was a 1-m diameter beryllium telescope used in the receiver. As ICESat orbited, the laser produced a series of approximately 70-m diameter laser spots that are separated by nearly 170-m along the spacecraft's ground track. Aerosol data could also be retrieved.



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 10**

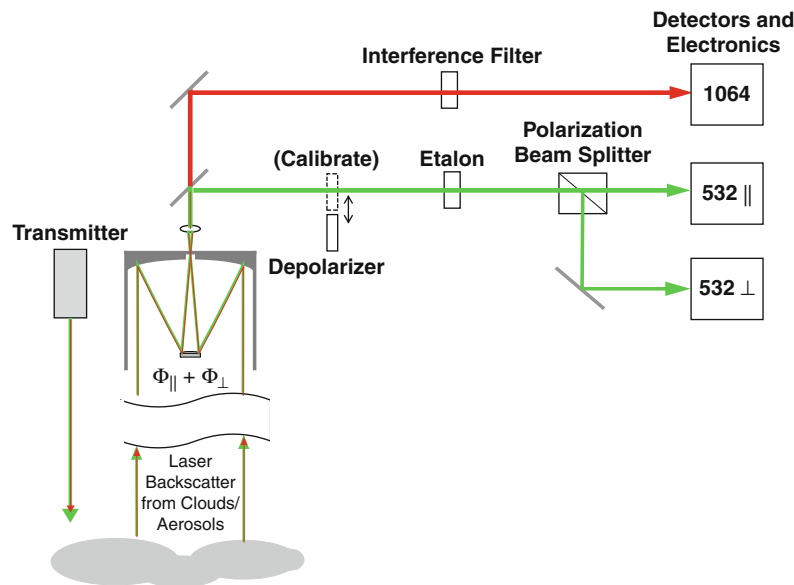
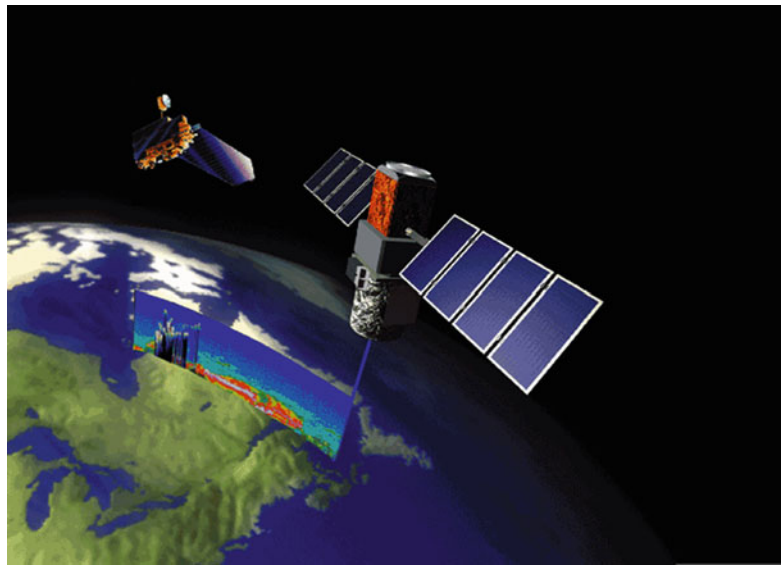
The analysis by which the DIAL raw data is transformed into useful data products

ICESat was a major step forward for space-based lidars. It lasted for some 5 years and used diode laser pumping of the Nd:YAG laser transmitter. This is the only technique that can achieve the electrical-to-optical conversion to make space lasers practical. It was also low voltage technology minimizing corona discharge due to high voltage. It demonstrated the flexibility of taking multiple science measurements such as ice sheet height, cloud heights, atmospheric boundary layer height, and aerosol concentrations. It could be operated autonomously in the radiation and vacuum environment of space for many years until the pump diodes and other components eventually degraded. All this capability is directly related to future DIAL systems deployed in space.

The only current operational lidar satellite in orbit is the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite

Observations (CALIPSO) carrying a backscatter lidar instrument as shown in Fig. 11 [18]. The satellite was launched in 2006 to an orbit altitude of 705 km and after 5 years is still operational, gathering global data on clouds and aerosols.

The Nd:YAG laser transmitter has two wavelengths 1,064 and 532 nm each transmitting 110 mJ, 20-ns linearly polarized laser pulses. The 532-nm channel can measure depolarization with a polarization beam splitter resolving parallel and perpendicular return signal components. The repetition rate is 20.16 Hz. The beams form a footprint of 70-m diameter with 335-m spacing between footprints. The receiver telescope is 1-m diameter with a field of view of 90 m on the surface. Aerosol profiles can be retrieved from 40-km altitude to the surface with 60-m vertical and 5-km horizontal resolutions.



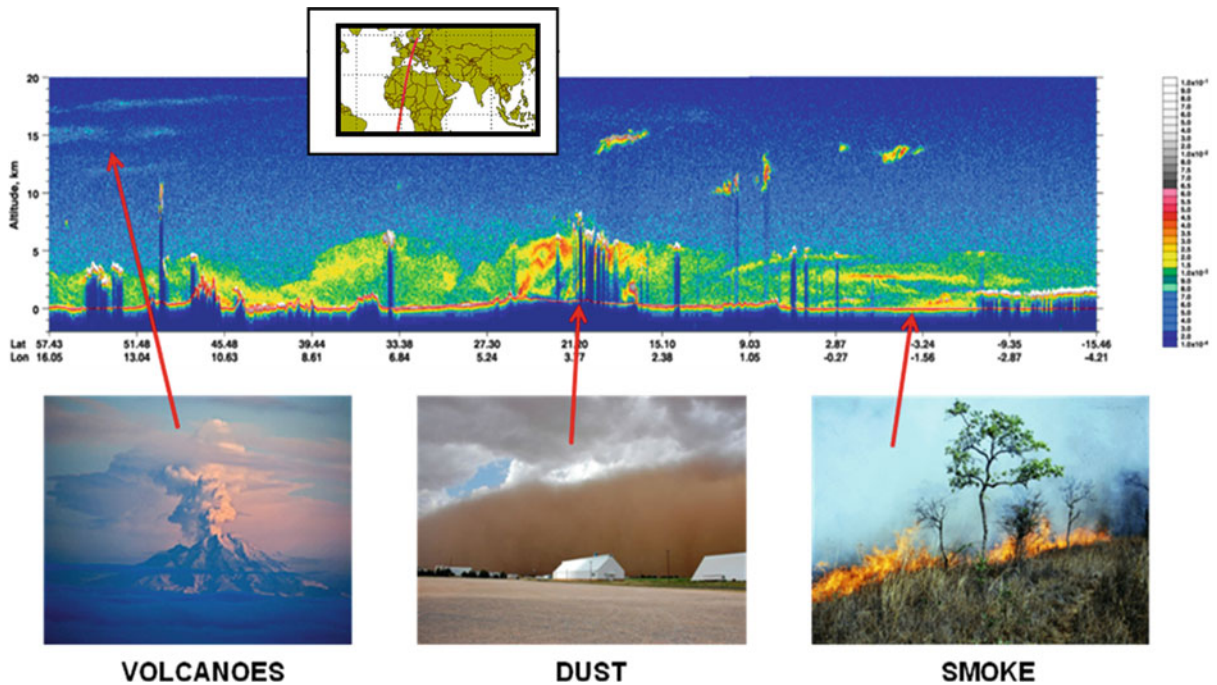
**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 11**

The CALIPSO satellite showing a typical cloud and aerosol lidar retrieval. Also a schematic of the lidar optical system

CALIPSO is providing the first multiyear global dataset of lidar aerosol and cloud profiles. Figure 12 is an example of one flight track on July 16, 2009 over the African Continent. The insert shows the orbit track from north to south. The aerosol and cloud 532-nm profile moves from left to right. CALIPSO has the ability to identify broad groups of aerosol types. In the upper left are aerosols emitted from volcanoes and in the center dust plumes are seen. Near the coast

of Africa, biomass burning is taking place and smoke plumes are noted near 3-km altitude. Over the ocean, clouds are forming at the top of the boundary layer. A space-based DIAL system would acquire aerosol and cloud data very similar to these CALIPSO results.

Two future lidar missions are planned by the European Space Agency the ADM-Aeolus [20] (launch 2013) and Earthcare [21] (launch 2016) satellites. The aim of the Atmospheric Dynamics Mission



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 12**  
 CALIPSO image shows 532-nm attenuated backscatter profiles from 16 July 2009 (~01:35 UTC until ~01:55 UTC) [19]

(ADM-Aeolus) is to provide global observations of three-dimensional wind fields, which will improve current wind-profiling atmospheric modeling, operational weather forecasting, and climate research. ADM-Aeolus satellite (400-km sun synchronous) will gather data by the active doppler wind lidar (DWL) method, whereby laser (Aladin-Atmospheric LAsER Doppler INstrument) pulses are transmitted in the atmosphere and then measure the backscattered Doppler shift of the received return signal, creating a wind profile showing the relative strength and direction of winds at different altitudes, as well as moisture and aerosol levels in the atmosphere. Wind speeds will be measured in the boundary layer (1-m/s resolution), in the free troposphere (2-m/s) with average wind velocities over 50-km tracks and 120 profiles per hour. The diode-pumped Nd:YAG laser is frequency tripled to 355 nm at 150-mJ pulse energy, 100-Hz pulse repetition rate, and 30-MHz laser linewidth (tunable over 10-MHz). The receiver has a Mie channel (aerosol and water droplets) with Fizeau spectrometer, and a Rayleigh channel (molecular scattering). The telescope is 1.5-m diameter.

The EarthCARE satellite (Earth Clouds, Aerosols, and Radiation Explorer) is a joint European-Japanese mission addressing the need for better understanding of the interactions between cloud, radiative and aerosol processes that play an important role in climate change by acquiring vertical profiles of clouds and aerosols, as well as the radiances at the top of the atmosphere. Cloud feedbacks are the main cause of uncertainty in predictions of future climate, and correct representation of clouds, aerosols, and radiation processes in models is needed. Current knowledge of the global profiles of aerosol and cloud properties is far too limited and the required profiles can only be provided by the high spectral resolution lidar. EarthCARE has been defined with the specific scientific objectives of quantifying aerosol-cloud-radiation interactions so they may be included correctly in climate and numerical weather forecasting models by providing:

- Vertical profiles of natural and anthropogenic aerosols on a global scale, their radiative properties and interaction with clouds.

- Vertical distribution of atmospheric liquid water and ice on a global scale, their transport by clouds and radiative impact.
- Cloud overlap in the vertical, cloud-precipitation interactions and the characteristics of vertical motion within clouds.
- The combination of the retrieved aerosols and cloud properties to derive the profile of atmospheric radiative heating and cooling.

The Atmospheric Lidar (ATLID) on the satellite is a high-spectral resolution with depolarization lidar system. The Nd:YAG laser produces 355-nm pulses at 30 mJ and a repetition rate of 74 Hz. High spectral purity demands a 50-MHz linewidth and stability of 50 MHz over a 1 month time in order to separate the Mie (100-m resolution) and Rayleigh (300 m) scattering atmospheric contributions by the high spectral resolution technique.

These backscatter lidar systems add to the technology base required by future space DIAL systems. The LITE system demonstrated that valuable aerosol and cloud data could be obtained from space. The ICESat mission used diode laser pumping, allowing efficient long-term lidar measurements from space. The CALIPSO mission is currently providing important long-term information on aerosols and clouds for climate models. In the future, ADM-Aeolus and ATLID EarthCARE missions will use large receiver telescopes and laser line narrowing techniques required to spectrally resolve atmospheric species by the DIAL technique for future missions.

While the DIAL technique is most often considered for measuring Earth's atmosphere, it could also be used for other planetary atmospheres such as Mars. A small lander could measure the water vapor concentration of the Mars atmosphere to help in the search for water-based life as life, as known on planet Earth, requires water. Although the atmospheric concentration of water vapor is very low, there is still sufficient vapor to form hazes, ground fog, frost, and clouds. An exchange of water vapor exists between the ice cap at the northern pole and the atmosphere. The DIAL technique could be utilized to measure the processes that determine the distribution of water vapor and thus indicate where sources and sinks of water vapor exist. This could then lead to sites where life on Mars might be found.

Figure 13 shows a Tm:germanate fiber laser, tunable around 1.9  $\mu\text{m}$ , which when coupled to a low mass lidar receiver, can fulfill the requirements for a water vapor Mars DIAL system. Laser diodes are used to pump a Te:germanate fiber that lases on a water vapor absorption line near 1.9  $\mu\text{m}$ . The DIAL system is very compact and low mass yet capable of measuring the water vapor profile up to about 6 km. Figure 14 shows the pulsed laser energy as a function of laser diode pumping current. The on- and off-line wavelengths are indicated where the online experiences more atmospheric absorption than the off-line wavelength. Such small systems could easily be packaged and deployed at multiple locations to profile the planetary water vapor distribution [22, 23].

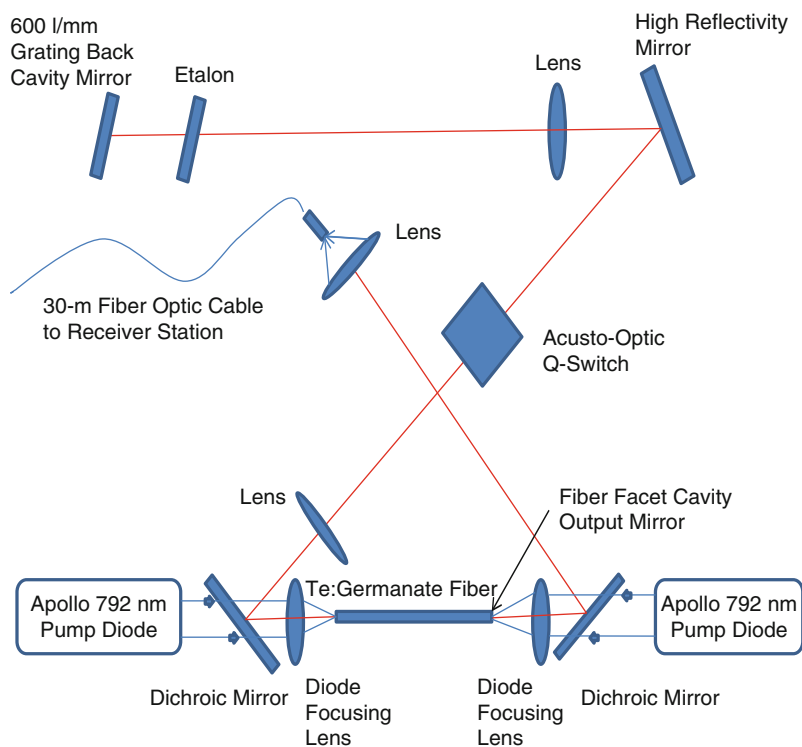
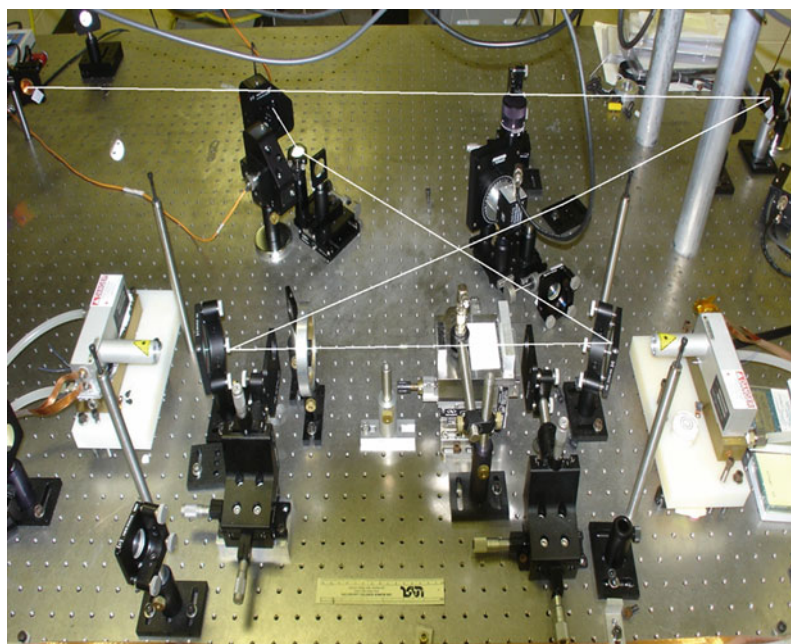
### Future Directions

Ground- and aircraft-based DIAL systems are now in regular use profiling many atmospheric species as well as clouds and aerosols [24]. The open question is when and how will the DIAL technique be deployed in space for global Earth and other solar system bodies atmospheric observations. The CALIPSO lidar system cost about \$300 million (FY 06 dollars) including launch, thus a DIAL system deployed in space would cost over \$1 billion due to its increased complexity when compared to a backscatter lidar such as CALIPSO.

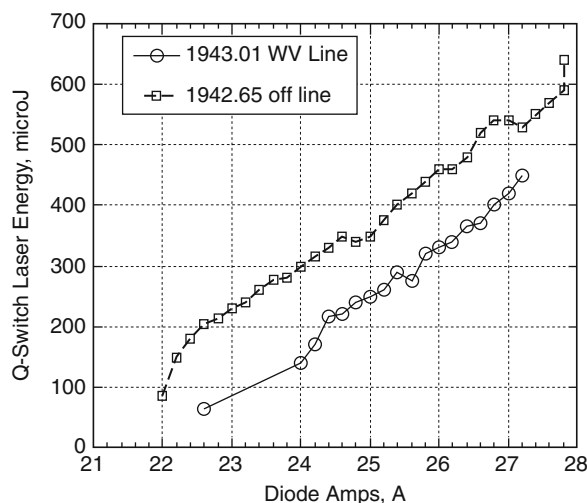
A number of studies have been published describing concepts and characteristics of space deployed DIAL systems for ozone [25] and water vapor [26, 27]. A concept for a space-based ozone DIAL system called ORACLE would deploy a UV DIAL system at 400 km with a laser pulse energy of 500 mJ and on- and off-line wavelengths of 308 and 320 nm, respectively. The ORACLE system would measure a single species, ozone, as well as clouds and aerosols. The receiver would have a 4.5  $\text{m}^2$  telescope. For daytime operation, the ozone concentration error would be  $\sim 10\%$  for altitudes between 5 and 45 km. Such studies demonstrated the utility as well as technology progress needed to deploy an ozone DIAL system in space.

With such a large investment in a space-based DIAL system, there would be a need to profile many different atmospheric species, some not recognized as important today for both air quality and global climate





**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL).** Figure 13 Schematic and photo of the Te:germanate fiber laser transmitter



**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 14**  
Pulsed laser output vs. pump diode current

change applications. In this concept, the DIAL system would rapidly tune to several different online wavelengths corresponding to the species of interest in the present atmospheric campaign. Thus the species involved in the major atmospheric chemical reactions would be profiled and then would be used to initialize global chemical weather models. The off-line wavelength would be used, as it is presently, to profile clouds and aerosols. The laser transmitter would need to be reliably tuned to any online wavelength with narrow linewidth and broad tuning range to cover all molecular species of interest.

The DIAL laser transmitter would need to operate in the infrared spectral region from about 1 to 5  $\mu\text{m}$  where there are many molecular absorption lines available for almost all atmospheric species of interest. The laser (orbit altitude 350–400 km) would need to generate 10-Hz pulse pairs at from 0.5 to 1.0-J energy per pulse. The receiver telescope would need to be about 5 m in diameter and the receiver optical system would need to be  $\sim 70\%$  efficient [28]. Avalanche photo detectors (APDs) such as the HgCdTe are being developed which can photon count the return signal. Such a DIAL system could typically profile atmospheric ozone with horizontal resolution of 200 km, a vertical resolution of 2 km with 10% measurement accuracy. This would

allow adequate profiling of regional and continental scale processes.

Future space-based DIAL systems will be driven by the science requirements that are related to the open questions with regard to climate change [29] and air quality [30, 31]. An example of how the above multiwavelength space-based DIAL system could impact an air quality global pollution field campaign can be demonstrated using the International Global Atmospheric Chemistry (IGAC) Lagrangian 2 K4 campaign [32]. In this field campaign, several aircraft were used to measure in situ concentrations of ozone, carbon monoxide, volatile organic compounds,  $\text{NO}_x$  ( $\text{NO} + \text{NO}_2$ ), and  $\text{NO}_y$  (mainly in the form of PAN) created from Alaskan forest fires in 2004 and transported to Europe. These fires create polluted plumes that undergo chemical evolution and mixing with air masses of different origin such that their chemical composition can radically change. Ozone levels in such polluted plumes are of interest since it is a major health concern and also a major source of the OH radicals (most efficient cleansing agent in the atmosphere). Ozone is mainly produced by reactions involving oxidation of volatile organic compounds (VOCs), and carbon monoxide (CO) in the presence of nitrogen oxides ( $\text{NO}_x = \text{NO} + \text{NO}_2$ ), and mainly destroyed by reactions involving water vapor and  $\text{HO}_x$  radicals. This long-range biomass burning plume was sampled in the troposphere over North America, the North Atlantic, and over Europe by three different aircraft. A photochemical trajectory model was used to examine the chemical processes responsible for the plume evolution. Six day modeled concentrations of  $\text{O}_3$ , CO,  $\text{NO}_y$ , NO, PAN,  $\text{H}_2\text{O}$  and temperature were compared to sparse in situ concentration measurements.

While campaigns such as the above are very costly and very infrequent, an orbiting IR tunable DIAL system could measure all the species giving a three-dimensional profile of the plume as it was transported toward Europe. The ability to accurately profile the chemical change as the plume aged demonstrates the unique and powerful ability to understand with high-resolution atmospheric chemical weather. Figure 15 shows the most important atmospheric molecular species to be profiled by DIAL and the corresponding need to measure these species [33–35].

| Molecular Species                                       | Measurement Need |                      |         |                              |
|---------------------------------------------------------|------------------|----------------------|---------|------------------------------|
|                                                         | Air quality      | Oxidation efficiency | Climate | Stratosphere ozone depletion |
| O <sub>3</sub> (ozone)                                  | ■                | ■                    | ■       | ■                            |
| CO (carbon monoxide)                                    | ■                | ■                    |         |                              |
| <i>j</i> (NO <sub>2</sub> )                             | ■                | ■                    |         |                              |
| <i>j</i> (O <sup>1</sup> D)                             | ■                | ■                    |         |                              |
| H <sub>2</sub> O (water vapour)                         | ■                | ■                    | ■       | ■                            |
| HCHO (formaldehyde)                                     | ■                | ■                    |         |                              |
| VOCs (volatile organic compounds)                       | ■                | ■                    |         |                              |
| Active nitrogen: NO <sub>x</sub> = NO + NO <sub>2</sub> | ■                | ■                    |         | ■                            |
| Reservoir species: HNO <sub>3</sub>                     | ■                | ■                    | ■       | ■                            |
| N <sub>2</sub> O                                        |                  |                      | ■       |                              |
| SO <sub>2</sub>                                         | ■                |                      | ■       |                              |
| Active halogens: BrO, ClO, OCIO                         |                  |                      |         | ■                            |
| Reservoir species: HCl, ClONO <sub>2</sub>              |                  |                      |         | ■                            |
| CO <sub>2</sub>                                         |                  |                      | ■       |                              |
| CH <sub>4</sub>                                         |                  | ■                    | ■       | ■                            |
| OH                                                      | ■                |                      |         |                              |

**Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar (DIAL). Figure 15** Important atmospheric molecular species that could be measured from space using DIAL and the corresponding measurement needed to profile these species

An IR tunable DIAL system could potentially measure all these species from space producing a three-dimensional profile of the chemical evolution and transport of these species with high spatial and temporal resolution.

As interest in transport of atmospheric pollution species and associated pollution species precursors increases, there may develop a need to monitor fluxes of these atmospheric species as they are transported across national boundaries. As noted above, the ADM-Aeolus satellite will have a lidar capable of measuring wind velocities as a function of altitude. This measurement can be combined with the IR DIAL space-based system resulting in the determination of molecular fluxes ( $\#/cm^2\cdot s$ ) across national boundaries for regulatory and health monitoring purposes.

Differential absorption lidar systems have been deployed on the ground in aircraft and eventually in space. The essential components of a space-based

DIAL system are now in place; efficient diode-pumped solid-state lasers, high pulse energy, narrow linewidth emission, large telescopes, autonomous operation, and radiation resistant long life operation. While more research needs to be invested in infrared tunable DIAL laser transmitters, nevertheless the ultimate utility of DIAL systems will be demonstrated in space deployed atmospheric measurement systems generating high spatial and temporal resolution atmospheric molecular profiles for chemical weather and health monitoring.

### Bibliography

1. Browell EV, Carter AF, Shipley ST, Allen RJ, Butler CF, Mayo MN, Siviter JH, Hall WM (1983) NASA multipurpose airborne DIAL system and measurements of ozone and aerosol profiles. *Appl Optics* 22:522–534
2. Browell EV, Browell EV (1983) In: Killinger DK, Moorradian A, Killinger DK, Moorradian A (eds) *Optical laser remote sensing*. Springer, New York, pp 138–148

3. Bucholtz A (1995) Rayleigh scattering calculations for the terrestrial atmosphere. *Appl Optics* 34:2765–2773
4. Williamson CK, De Young RJ (2000) Method for the reduction of signal-induced noise in photomultiplier tubes. *Appl Optics* 39:1973–1979
5. Weitkamp C (2005) Lidar range resolved optical remote sensing of the atmosphere. Springer, New York
6. Kovalev VA, Eichinger WE (2004) Elastic lidar. Wiley, New Jersey
7. Browell EV (1989) Differential absorption lidar sensing of ozone. *Proc IEEE* 77:419–432
8. Kuang S, Burris JF, Newchurch MJ, Johnson S, Long S (2010) Differential absorption lidar to measure subhourly variation of tropospheric ozone profiles. *IEEE Trans Geo Remote Sensing* 49:557–571
9. <http://www.espo.nasa.gov/arctas/>
10. Flesia C, Mugnai A, Emery Y, Godin L, de Schouepnikoff L, Mitev V (1994) Interpretation of lidar depolarization measurements of the Pinatubo stratospheric aerosol layer during EASOE. *Geo Res Letts* 21:1443–1446
11. Schouepnikoff L, Van Den Bergh H, Calpini B (1998) Tropospheric air pollution monitoring LIDAR. In: Myers RA (ed) *Encyclopedia of environmental analysis and remediation*. Wiley, New York, pp 4873–4909
12. Ismail S, Browell EV (1994) Recent lidar technology developments and their influence on measurements of tropospheric water vapor. *J Atmos Oceanic Tech* 11:76–84
13. Ismail S (2011) GRIP Campaign data, private communication
14. Kooi S (2011) private communication
15. Winker DM, Couch RH, McCormick MP (1996) An overview of LITE: NASA lidar in-space technology experiment. *Proc IEEE* 84:164–179
16. Farrell SL, Laxon SW, McAdoo DC, Yi D, Zwally HJ (2009) Five years of arctic sea ice freeboard measurements from the ice, cloud and land elevation satellite. *J Geophys Res* 114:C04008. doi:10.1029/2008JC005074
17. Abdalati W, Zwally HJ, Bindschadler R, Csatho B, Farrell SL, Fricker HA, Harding D, Kwok R, Lefsky M, Markus T, Marshak A, Neumann T, Palm S, Schutz B, Smith B, Spinhrne J, Webb C (2010) The ICESat-2 laser altimetry mission. *Proc IEEE* 98:735–751
18. Winker DM, Pelon J, Coakley JA, Ackerman SA, Charlson RJ, Colarco PR, Flamant P, Fu Q, Hoff RH, Kittaka C, Kubar TL, Le Treut H, McCormick MP, Megie G, Poole L, Powell K, Trepte C, Vaughan MA, Wielicki BA (2010) The CALIPSO mission a 3D view of aerosols and clouds. *Bull Am Met Soc*. doi:10.1175/2010BAMS3009.1
19. Vaughan MA (2011) NASA Langley Research Center, private communication
20. Clissold P (ed) (2008) ADM-Aeolus, SP-1311, ESA Communication Production Office, The Netherlands, ISBN 978-92-9221-404-3
21. Le Hors L, Toulemont Y, Heliere A (2008) Design and development of the backscatter lidar ATLID for EARTHCARE. In: *International Conference on Space Optics*, Toulouse, 14–17 Oct 2008
22. Barnes NP, Walsh BM, Reichle D, De Young RJ (2009) Tm: fiber lasers for remote sensing. *Opt Mat*. doi:10.1016/j.optmat.2007.11.037, 31:1061–1064
23. De Young RJ, Barnes NP (2010) Profiling atmospheric water vapor using a fiber laser lidar system. *Appl Optics* 49:562–567
24. Browell EV (1995) Airborne lidar measurements. *Rev Laser Eng* 23:135–141
25. Ball DJ, Dudenzak AE, Rheault F, Browell EV, Ismail S, Stadler JH, Hoff RM, McElroy CT, Allan I, Carswell CTA, Hahn JF, Ulitsky A (1998) ORACLE (ozone research with advanced cooperative lidar experiment): joint NASA-CSA development of a space-base ozone DIAL. *Proc SPIE* 3494:223–226. doi:10.1117/12.332422
26. Ismail S, Browell EV (1989) Airborne and spaceborne lidar measurements of water vapor profiles: a sensitivity analysis. *Appl Optics* 28:3603–3615
27. Wulfmeyer V, Bauer H, Girolammo D, Serio C (2005) Comparison of active and passive water vapor remote sensing from space: an analysis based on the simulated performance of IASI and space borne differential absorption lidar. *Remote Sens Environ* 95:211–230
28. Ismail S, Peterson LD, Hinkle JD (2005) Applications of deployable telescopes for earth-observing lidar. In: *Proceedings of the ESTO earth science technology conference*, College Park, 27–28 June 2005
29. Chin M, Kahn RA, Schwartz SE (eds) (2009) CCSP 2009: atmospheric aerosol properties and climate impacts, Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, National Aeronautics and Space Administration, Washington, DC, 128 pp
30. Monks PS, Granier C, Fuzzi S, Stohl A, Williams ML, Akimoto H, Amanni M et al (2009) Atmospheric composition change – global and regional air quality. *Atmos Environ* 43:5268–5350
31. Browell EV, Ismail S, Grant WB (1989) Differential absorption lidar (DIAL) measurements from air and space. *Appl Phys B* 67:399–410
32. Real E, Law KS, Weinzierl B, Fiebig M, Petzold A, Wild O, Methven J, Arnold S, Stohl A, Huntrieser H, Roiger A, Schlager H, Stewart D, Avery M, Sachse G, Browell E, Ferrare R, Blake D (2007) Processes influencing ozone levels in Alaskan forest fire plumes during long-range transport over the North Atlantic. *J Geo Res* 112:D10S41. doi:10.1029/2006JD007576
33. Laj P, Klausen J, Bilde M, Pla-Duelmer C, Pappalardo G, Clerbaux C, Baltensperger U et al (2009) Measuring atmospheric composition change. *Atmos Envir* 43: 5351–5414
34. Ismail S, Gervin J, Wood HJ, Peri F (2004) Remote sensing of tropospheric chemistry using lidars from geostationary orbit. *Proc SPIE* 5659:146–154
35. World Meteorological Organization (2007) WMO global atmospheric watch strategic plan: 2008–2015. Report 172

## Alaska Gas Hydrate Research and Field Studies

S. L. PATIL, A. Y. DANDEKAR, S. KHATANIAR

Institute of Northern Engineering, University of Alaska  
Fairbanks, Fairbanks, AK, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Global Distribution of Hydrate Resource

Gas Hydrate Resource Potential

Review of Global Gas Hydrate Projects

History of Gas Hydrate Exploration in Alaska

Alaska Gas Hydrate Production Potential

North Slope Borough Methane Hydrate Project Review

Future Directions

Conclusions

Disclaimer

Acknowledgment

Bibliography

### Glossary

**In-place** The term used to estimate gas hydrate resources disregarding technical or economical recoverability. Generally these are the largest estimates [1].

**Permafrost** A layer of soil or bedrock that has been continuously frozen for at least 2 years.

**Petrophysical** Physical properties of the reservoir rock.

**Resources** Resources are those detected quantities of hydrocarbon that cannot be recovered profitably with the current technology but may be recoverable in the future. It also includes those quantities that geologically may be possible but have not been found yet [2].

### Definition of the Subject

#### Gas Hydrates

Natural gas hydrates are solid, crystalline, ice-like materials made up of small molecules of gases, mainly

methane confined inside cages of water molecules. These gas hydrates are solid solutions and are formed when water molecules linked by hydrogen bonding create cavities (host) that enclose a variety of molecules (guest). These molecules can be  $\text{CH}_4$ ,  $\text{C}_2\text{H}_6$ , and  $\text{CO}_2$  to name a few. It is interesting to note that there is no chemical bonding between the host water molecule and the caged molecule. Hydrates can form under high pressures or low temperatures, which are generally the conditions that exist in the Arctic or under deep ocean beds. Huge volumes of gas hydrates are expected to exist at various locations around the globe (Fig. 3). It is known that  $1 \text{ m}^3$  of hydrates upon dissociation release about  $180 \text{ std m}^3$  of gas, making them a huge potential as a future energy resource [3, 4].

### History of Gas Hydrates

The discovery of hydrates dates way back in 1810 when Sir Humphry Davy conducted an experiment and discovered that an aqueous solution of chlorine can be converted into solid form when cooled below  $9^\circ\text{C}$  [5]. Faraday confirmed the existence of this solid compound and now it is known that there are more than 100 molecules that can form hydrates when combined with water. Few molecules that are very commonly found to form hydrates are  $\text{CH}_4$ ,  $\text{C}_2\text{H}_6$ ,  $\text{C}_3\text{H}_8$ ,  $\text{CO}_2$ , and  $\text{H}_2\text{S}$ , with  $\text{CH}_4$  being the most common molecule to form gas hydrates found in existence. It was as early as the 1950s when through x-ray diffraction methods the crystal structure of hydrates was known, and by late 1960 it was realized that the earth's crust is home to vast quantities of natural gas hydrates [5].

It was Powell, in 1948 at the University of Oxford, who was the first to describe clathrate structure and described clathrate compounds as those inclusion compounds *in which two or more components are associated without ordinary chemical union but through complete enclosure of one set of molecules in a suitable structure formed by another* [5]. The clathrate compounds are divided into two categories. One is aqueous clathrates in which water is the host species, commonly known as clathrate hydrates or gas hydrates, and the other category of hydrates where water is not the host molecule are known as nonaqueous clathrates [5]. It is known that water molecules are linked to each other through hydrogen bonding, which results in the



formation of cavities having dimensions of 780 and 920 pm [5]. The molecules that are smaller than the dimension of these small cavities and do not react with the hydrogen bonds of the water molecule form gas hydrates under appropriate conditions of temperature and pressure, thereby providing stability to the structure.

## Introduction

Gas hydrates may become an important global source of clean burning natural gas. Scientists have considered the potential of gas hydrate as an energy resource for nearly four decades. However, until 2002, this knowledge has not been systematically applied to determine the economic practicality of gas hydrate resource development. The estimates of in-place unconventional gas hydrate volume are up to 100 Trillion Cubic Feet (TCF) within the existing oil development infrastructure of the Alaska North Slope (ANS) [6]. This region uniquely combines known gas hydrate presence, an undetermined volume of associated free gas accumulations, and existing production infrastructure (Fig. 1).

Initial research on Alaskan gas hydrates was focused on the Prudhoe Bay Unit (PBU), Kuparuk River Unit (KRU), and Milne Point Unit (MPU) infrastructure areas of the Alaska North Slope. These areas were being studied to determine gas hydrate reservoir extent, stratigraphy, structure, continuity, quality, variability, and geophysical and petrophysical property distribution. These initial studies characterized reservoirs and fluids, which led to in-place and recoverable resource estimates and commercial potential determination. Studies were conducted to determine best practices and procedures for gas hydrate drilling, data acquisition, completion, and production. This collaborative research provided practical input to reservoir and economic models, determined the technical feasibility of gas hydrate production, and influenced future exploration and field extension of this potential unconventional gas resource.

A successful drilling project to explore, confirm, and evaluate gas hydrates was conducted in February 2007 through a shallow seismic anomaly interpreted as hydrates. In this project, over 400 ft of sediments containing gas- and water-bearing core and well-developed gas hydrate zones (85% recovery) were cored. An extensive suite of electric logs were run to

identify the gas hydrate zones, and gas hydrate saturations were computed. Open-hole pressure-transient analyses and fluid sampling tests were conducted using Schlumberger's Modular Formation Dynamics Tester (MDT™).

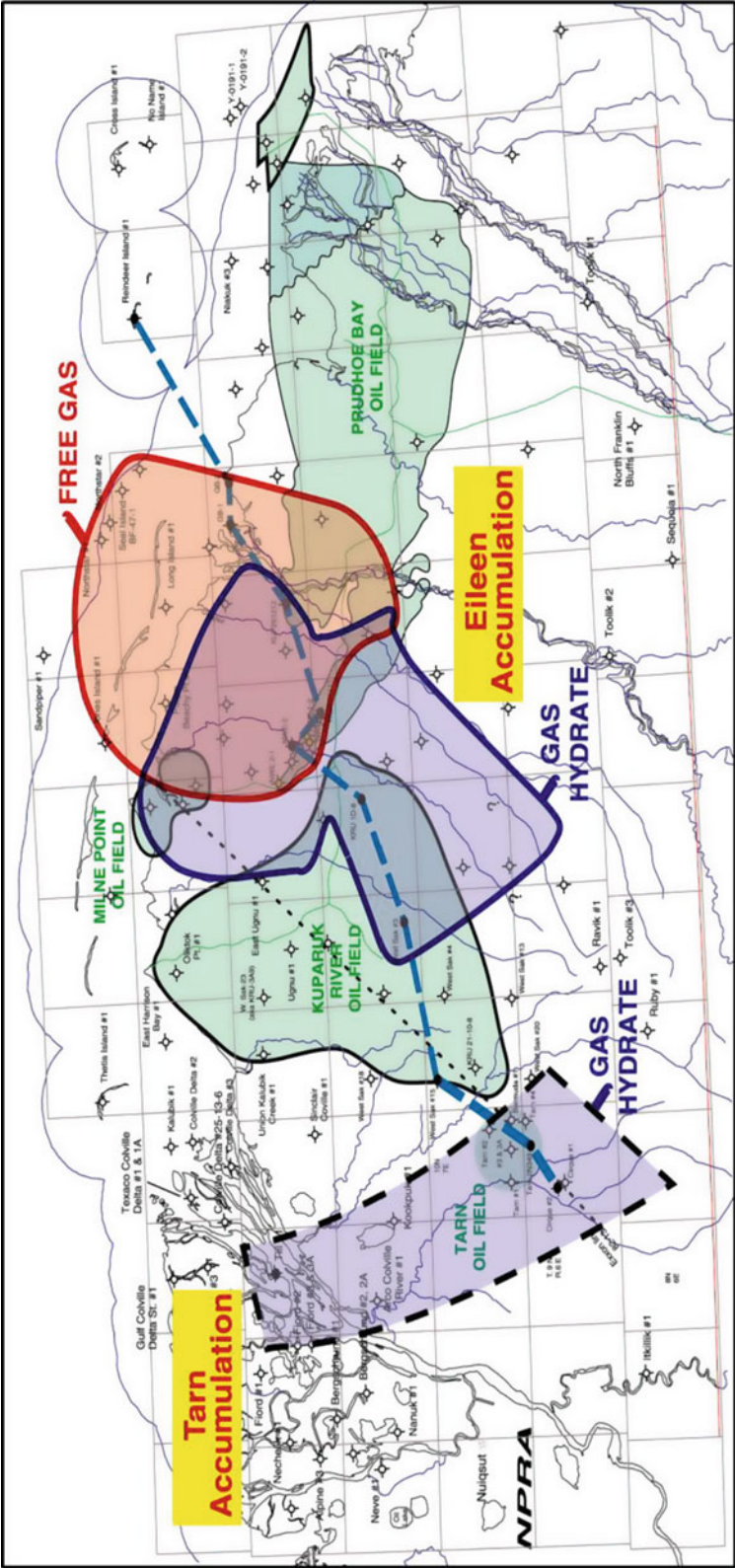
A second ANS gas hydrate research project was conducted for the Barrow Gas Field (BGF). Barrow, Alaska, is a remote and the northernmost city in the United States. The energy demands of the city and nearby localities are met by natural gas supply from the Barrow Gas Field consisting of three different gas pools: the East Barrow (EB) pool, the South Barrow (SB) pool, and the Walakpa (WAL) pool (Fig. 2).

Gas hydrate resource potential of the three BGF pools was analyzed by developing gas hydrate stability models. Material balance studies were performed on East Barrow (EB) gas pool to understand reservoir drive mechanisms and qualitatively estimate the hydrate resource. Field-scale dynamic reservoir simulation models were developed for the EB and WAL gas pools. Production history data were matched, reservoir drive mechanisms were confirmed, free gas and hydrate resources were quantified, hydrate dissociation patterns were examined, optimum locations for drilling infill wells were identified, and future production scenarios were simulated.

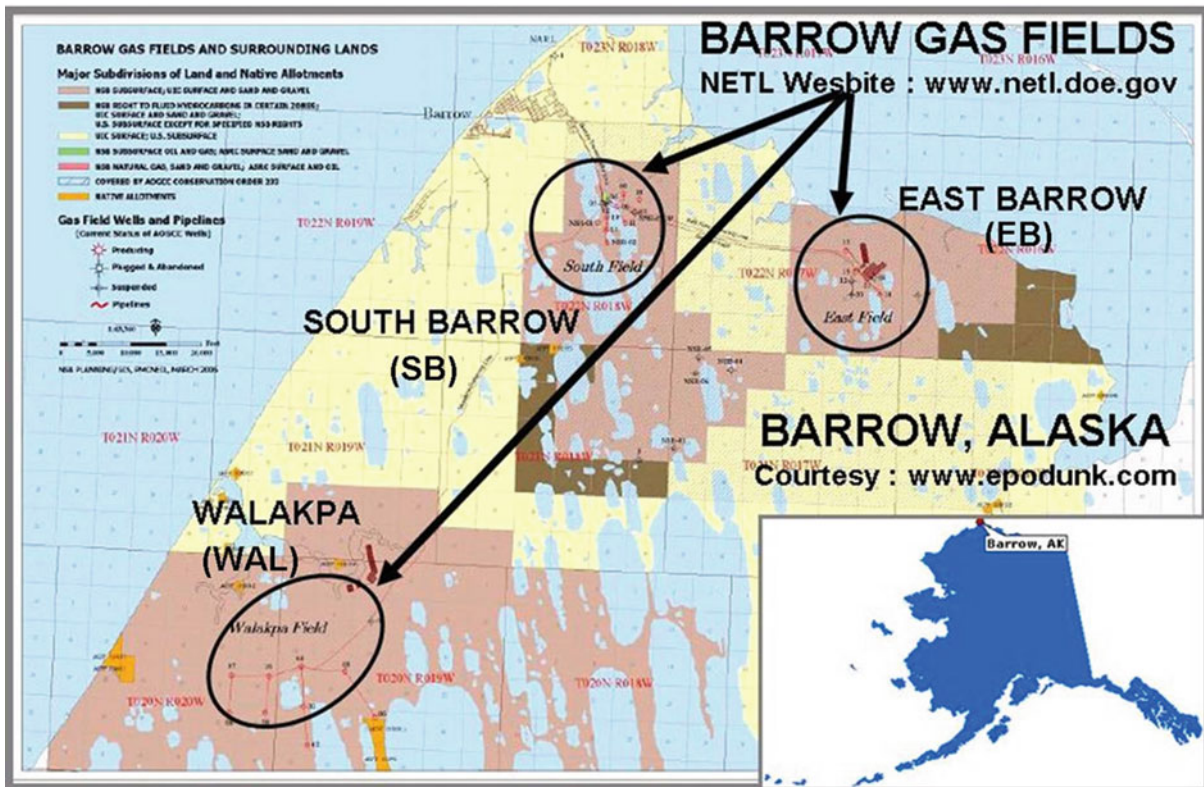
The vast amount of unconventional gas in-place combined with evaluation of conventional Alaska North Slope gas commercialization creates industry-government alignment to assess this potential future resource. This region uniquely combines known gas hydrate presence, an undetermined volume of associated free gas accumulations, and existing production infrastructure. However, many technical and economic challenges must be resolved before the gas hydrate becomes a viable commercial hydrocarbon resource.

## Global Distribution of Hydrate Resource

Global volumes of gas hydrate estimated till now span several orders of magnitude and there is no certainty in methane hydrate resource estimates across the globe. This uncertainty in estimating the volume of gas hydrates present is majorly because of two factors: (1) The commonly used seismic methods and other geophysical methods that are used in estimating conventional hydrocarbon resources are not sensitive to the huge volumes of hydrates present in the seafloor



Alaska Gas Hydrate Research and Field Studies. Figure 1  
Alaska North Slope oil field developments with Eileen and Tarn gas hydrate trends (Courtesy Collett [32])



Alaska Gas Hydrate Research and Field Studies. Figure 2

Barrow gas fields, Alaska (Modified from NETL Website, [www.netl.doe.gov](http://www.netl.doe.gov))

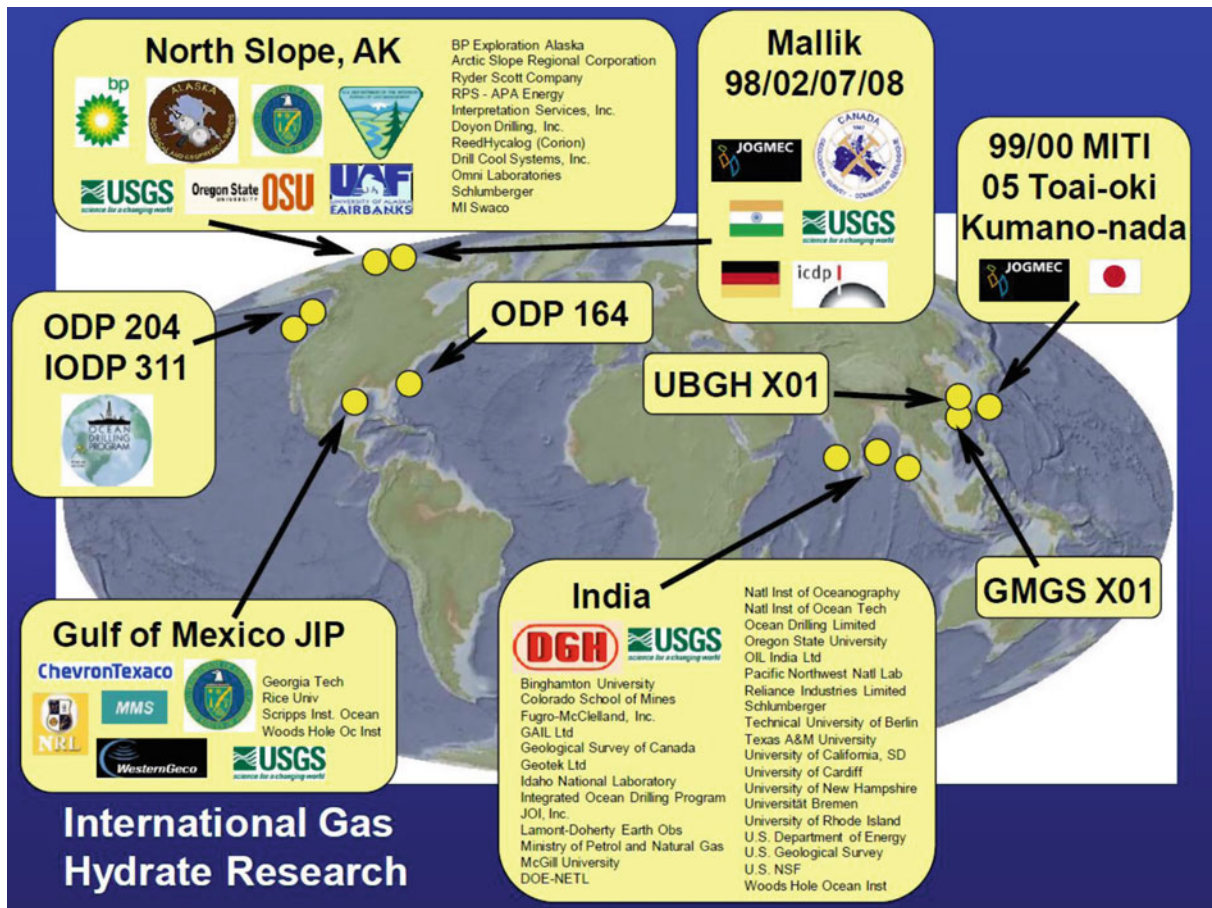
section. (2) Drilling activities to directly confirm the presence of hydrates have a limited applicability due to limited spatial extent [7]. There is no direct evidence which can be used in estimating the occurrence of in-place gas hydrates. There are 23 locations where the presence of gas hydrates has been confirmed through the recovery of hydrate samples, 20 of them being in the oceanic and the rest 3 in the arctic environments [8, 9]. Since the discovery of gas hydrates, researchers have tried to estimate the amount of gas trapped in hydrates none being certain. Going by even the most conservative approximation of the estimation of the trapped gas, the amount of gas trapped in hydrates is enormous.

HEI (Hydrate Energy International) along with the International Institute for Applied Systems Analysis (IIASA) with funding from the World Bank, United Nations Organizations, and national governments has undertaken the task of assessing the global gas hydrate potential. For this study, data from every continental margin was studied and appropriate depositional

models were utilized. The data was reported for 18 regions defined by the United Nations. For the Arctic Ocean, separate resource assessment was done without regard for national boundaries and for the southern ocean (from the coast of Antarctica north to 60° south latitude) [10]. The result of global hydrate resource estimate has been tabulated in Table 1 after Johnson [10]. Figure 4 depicts the hydrate resource pyramid after Boswell and Collett [11].

The production and commercialization of gas hydrate deposits rely heavily on two important factors: (1) the advancement in technology to prove and produce from these resources overcoming existing obstacles and (2) a natural gas supply market accelerating the need to progress and improve the gas hydrates resource potential [12]. Although various research activities are on way worldwide, gas hydrates are the most unlikely unconventional resource that will be tapped in the near future. But based on the demand for energy in the future, these research and exploration activities need





Alaska Gas Hydrate Research and Field Studies. Figure 3

World distribution of gas hydrates (Adapted from oral presentation at AAPG Annual Convention, San Antonio, Texas, April 20–23, 2008)

to go on and probably in a much more aggressive way to make these a viable energy source. Figure 5 depicts the time line for post 1990 global drilling efforts after Ruppel [12].

### Gas Hydrate Resource Potential

Gas hydrates are naturally occurring ice-like substances composed of water and gas, in which a solid water-lattice accommodates gas molecules in a cage-like structure called a clathrate. This clathrate structure can store up to 180 unit volumes of methane gas per 1 unit volume clathrate. This large gas storage capacity makes methane hydrate a very attractive concentrated potential unconventional resource.

Gas hydrates accumulate within reservoirs located in permafrost regions and within subsea sediments of outer continental margins. While methane, propane, and other gases can be included in the hydrate structure, methane hydrates appear to be the most common in nature. The amount of methane sequestered in gas hydrates is probably enormous. However, the in-place and recoverable gas estimates are highly speculative and also dependent upon the conventional petroleum system components of source, charge, reservoir, trap, seal, and a highly uncertain recovery factor, as well as occurrence within the gas hydrate stability field (Fig. 6).

The successful production history of the Russian Messoyakha gas hydrate field [6] demonstrates that gas hydrates may provide an immediate source of natural

**Alaska Gas Hydrate Research and Field Studies. Table 1**  
Gas in-place in hydrate-bearing sands (After Johnson [10])

| Region (United Nations designation) | Gas in-place range (TCF) | Gas in-place median (TCF) |
|-------------------------------------|--------------------------|---------------------------|
| United States                       | 1,500–15,434             | 7,013                     |
| Canada                              | 533–8,979                | 2,228                     |
| Western Europe                      | 36–14,858                | 1,425                     |
| Central and Eastern Europe          | 0–105                    | 13                        |
| Former Soviet Union                 | 1,524–10,235             | 3,829                     |
| North Africa                        | 6–1,829                  | 218                       |
| Eastern Africa                      | 42–25,695                | 1,827                     |
| Western and Central Africa          | 79–26,672                | 3,181                     |
| Southern Africa                     | 121–26,369               | 3,139                     |
| Middle East                         | 31–3,848                 | 573                       |
| China                               | 10–1,788                 | 177                       |
| Other East Asia                     | 14–2,703                 | 371                       |
| India                               | 36–6,268                 | 933                       |
| Other South Asia                    | 20–3,497                 | 557                       |
| Japan                               | 71–471                   | 212                       |
| Oceania                             | 38–6,750                 | 811                       |
| Other Pacific Asia                  | 64–25,946                | 1,654                     |
| Latin America and the Caribbean     | 258–31,804               | 4,940                     |
| Southern Ocean                      | 144–45,217               | 3,589                     |
| Arctic Ocean                        | 178–55,524               | 6,621                     |
| Total                               | 4,705–31,3992            | 43,311                    |

gas that can be produced by conventional depressurization of an adjacent free gas. Gas hydrates also represent a significant drilling and production hazard. Russian, Canadian, and American researchers have described numerous problems associated with gas hydrates, including well control problems and casing failures.

### Review of Global Gas Hydrate Projects

Several countries have active gas hydrate field programs (Fig. 3). The following gives an overview/status of these research activities.

### Japan

The first remarkable gas hydrate research program in Japan was established in 1995 by the Government of Japan along with the Ministry of Economy, Trade, and Industry (METI). Since then the METI was very proactive in gas hydrate research field. It contributed to Canada's Mallik 2002 Gas Hydrate Production Research Well Program. In 2001, METI established Methane Hydrate R&D Program, which is currently in its Phase 2 (2009–2015). Within the Phase 2 production tests on the eastern Nankai Trough are planned for 2012 and 2014 respectively. Phase 2 also considers environmental impact assessment and development of optimal methane hydrate production approaches as its objectives [13].

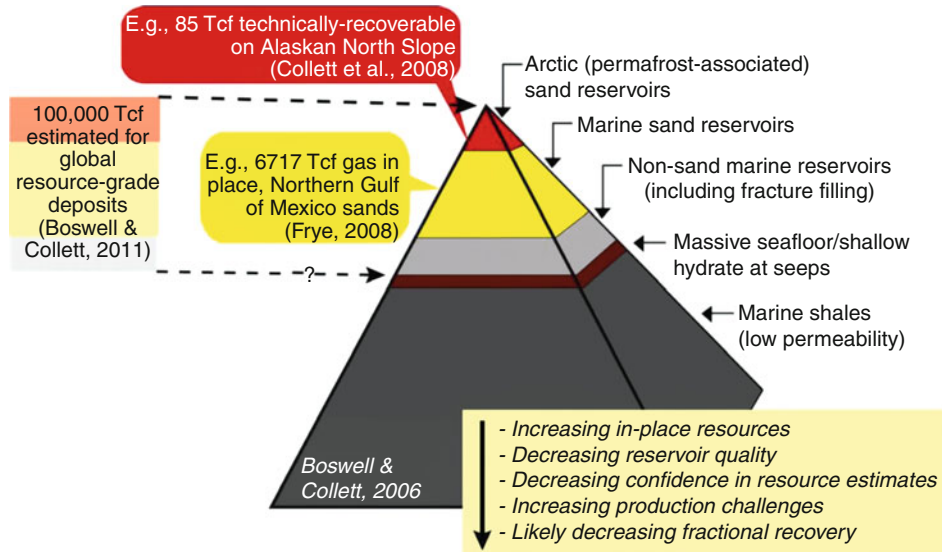
### China

Guangzhou Center for Gas Hydrate research was founded in 2004 to expand the research and to evaluate gas hydrate resource potential in offshore China. Guangzhou Marine Geological Survey (GMGS) conducted and expedition GMGS-1 in spring 2007 to perform offshore drilling in Shenhu area of the South China Sea. During the expedition eight holes were drilled. Wireline logging was performed on each site. Five wells were cored [14]. Future expeditions to the South China Sea are being considered. Besides, hydrate exploratory drilling takes place in Qinghai-Tibet Plateau.

### South Korea

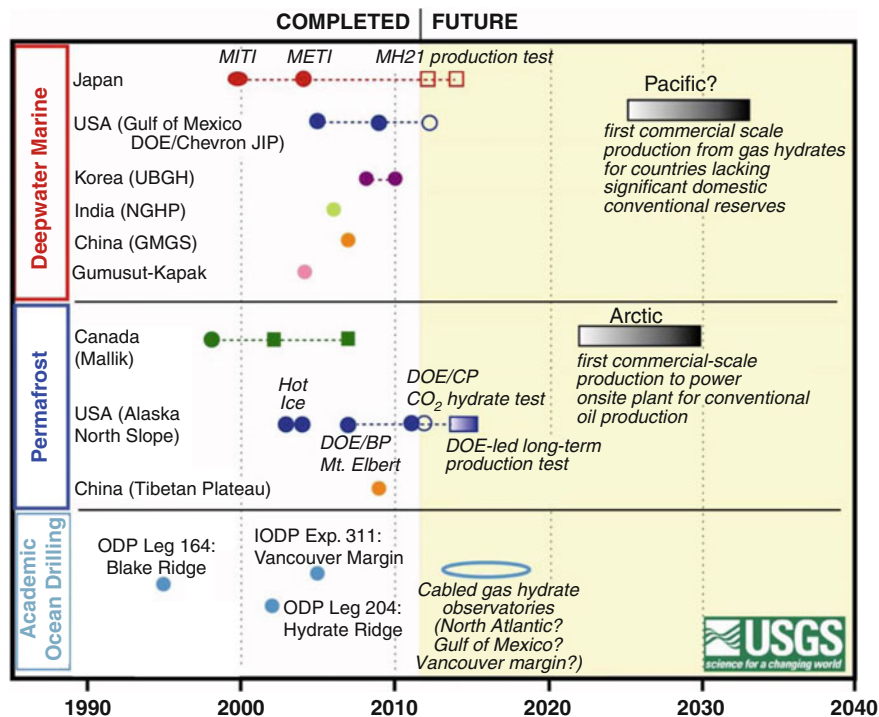
Gas Hydrate Research and Development Organization (GHDO-K), supported by the Ministry of Commerce, Industry, and Energy, is a South Korean gas hydrate program started in 2000. Its main objectives are to confirm gas hydrate resources in the East Sea and to commercially produce gas hydrates by 2015 [14]. It started with seismic data acquisition in the East Sea with shallow coring in the Ulleung Basin in order to define its gas hydrate potential. Ulleung Basin Gas Hydrate Expedition 1 (UBGH1) was the first exploration and drilling expedition launched on behalf of the Korean Institute of Geoscience and Mineral Resources in late 2007. During the expedition five wells were drilled with logging, and while drilling, some of them





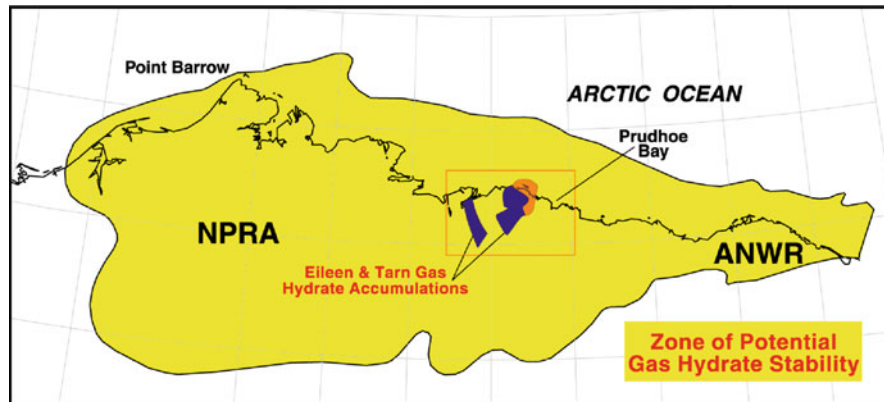
Alaska Gas Hydrate Research and Field Studies. Figure 4

The hydrate resource pyramid (Ruppel [12])



Alaska Gas Hydrate Research and Field Studies. Figure 5

Timeline of major post-1990 gas hydrate field programs and future activities. Circles correspond to logging and/or coring, while squares/rectangles denote activities that included/will include production testing. Solid symbols denote completed activities, and open symbols are potential or planned activities. Rectangles filled with shaded pattern refer to longer-term production activities that remain prospective (After Ruppel [12])



**Alaska Gas Hydrate Research and Field Studies. Figure 6**  
Zone of potential gas hydrate stability, Alaska, North Slope (Courtesy Collett [32])

were cored as well [14]. In July 2010, the UBGH2 expedition was launched. It included vast amount of data acquired and lasted till early October. Subsequently, South Korea is planning field production tests (GHDO-K, 2010).

### India

National Gas Hydrate Program (NGHP) of India was initiated in 1997. Its primary goals are performance of ocean drilling, coring, logging, and other activities in order to explore gas hydrate deposits and evaluation of the possibility of their economical production. The NGHP Expedition 1 was successfully conducted during mid-2006. During the expedition, 39 wells were drilled in various sites which are: Kerala–Konkan, Krishna–Godavari, Mahanadi, and Andaman. Special attention was paid to Krishna–Godavari gas hydrate site, where more than 10 wells were drilled [14]. Extensive coring and logging were performed in a number of wells and vast amounts of data were collected. NGHP Expedition 02 was also considered as a future possibility.

### United States

In the United States, the Department of Energy (DOE) has been conducting extensive research on gas hydrates for the past several decades. Ocean Drilling Program (ODP) Leg 164 located on Blake Ridge, Atlantic shelf, was the first special attempt to identify gas hydrates [15]. It recovered hydrate cores and documented gas hydrate presence within the sediment [14]. ODP Leg

204 was another deep-sea expedition which discovered gas hydrates in the Hydrate Ridge, located on the Cascadia continental margin. Gulf of Mexico Gas-hydrate JIP consortium was formed in 2001 by industry participants and government agencies in partnership with DOE and led by Chevron. Its goals were studying hydrates both as drilling hazards and also as a potential methane resource. In 2005, it conducted scientific drilling and coring in the gulf [14]. On the North Slope of Alaska, gas hydrates encountered onshore are associated with permafrost. Despite an apparent likelihood of offshore gas hydrate existence in the Arctic Ocean, there is no proof so far. In 2002–2003 Anadarko Petroleum Corporation drilled the Hot Ice gas hydrate research well in Tarn gas hydrate accumulation on the North Slope of Alaska. However, no gas hydrates were encountered [14]. In 2007, the Mount Elbert Gas Hydrate Stratigraphic Test Well was drilled by BP Exploration (Alaska) on behalf of DOE. It was located on the Milne Point, Eileen gas hydrate accumulation. It revealed gas hydrates and even cores were recovered.

### Canada

Canada does not have any official gas hydrate program. However, extensive study was performed in Canada of which the remarkable ones are Cascadia and Mallik. Both gas hydrate occurrences are associated with permafrost. ODP Leg 146 and IODP Expedition 311 to the offshore Vancouver Island discovered gas hydrate existence in Cascadia continental margin. Numerous

data including seismic, logs, and chloride concentration infer gas hydrate presence. Within IODP Expedition 311, hydrate-bearing cores were retrieved. Mallik-38 onshore well in the Mackenzie River Delta was believed to be drilled through 100 m of hydrate-bearing sandstone. Subsequently in 1998, Geological Survey of Canada on behalf of Japan Petroleum Exploration Company Ltd and Japan National Oil Corporation drilled Mallik 2L-38 gas hydrate research well. Coring and logging both revealed gas hydrates within the sandstone. As a next step, within Mallik 2002 program the partners drilled three more wells: Mallik 3L-38, Mallik 4L-38, and Mallik 5L-38 [14].

### New Zealand

In 2010, Gas Hydrate Resources (GHR) program was funded for 2 years by the New Zealand Foundation for Research, Science, and Technology (FRST). It is mostly focused on Hikurangi margin located to the east of the islands. The program plans are: analysis of existing 2D and 3D seismic data acquisition, seafloor sampling in spring 2011, and production modeling [16].

### History of Gas Hydrate Exploration in Alaska

The North Slope of Alaska contains two large gas hydrate accumulations within and near the Prudhoe Bay and the Tarn oil field developments (Fig. 1). This confirms the possibility that gas hydrates may represent an important future energy resource. The occurrence of gas hydrates on the North Slope of Alaska was confirmed in 1972 with data from the Northwest Eileen State-2 well, the first dedicated gas hydrate core and test well, located near the Prudhoe Bay Oil Field. Most of the gas hydrates near this Eileen trend discovery well occur in a series of laterally continuous sandstone reservoirs and are geographically restricted to the area overlying the western part of the Prudhoe Bay Oil Field and the eastern part of the Kuparuk River Oil Field. Three-dimensional seismic surveys in the Prudhoe Bay Oil Field indicate the presence of an undetermined amount of free gas trapped down deep below some of the gas-hydrate-bearing units. The volume of gas within the gas hydrates within the Eileen trend is estimated to be about 37–44 trillion cubic feet.

Recently, data from wells along the western margin of the Kuparuk River Field have shown presence of the

large Tarn trend gas hydrate accumulation overlying the Tarn oil field. The Cirque-1 well, located about 4 miles southwest of the Tarn oil field, experienced severe well control problems in 1992 after drilling through what appeared to be a free gas interval possibly trapped below the gas hydrate stability zone. The gas-hydrate-bearing stratigraphic interval within the Tarn trend appears to be the up-dip equivalent of the West Sak sands. These sands are estimated to contain more than 20 billion barrels of in-place heavy oil and are the focus of recent development activity. Preliminary analyses of other recently completed wells along the western margin of the Kuparuk River oil field suggest that the volume of in-place gas hydrates within the Tarn trend may exceed that of the Eileen trend.

### Alaska Gas Hydrate Production Potential

The production potential of the Eileen or Tarn gas hydrate accumulations has not been adequately tested. This area is a focus of the gas hydrate research programs by the industry, led by the US Department of Energy. In order to produce gas from gas hydrate, the in situ pressure-temperature regime must change to dissociate gas. Decreasing the reservoir pressure, increasing the reservoir temperature, and/or use of chemicals can allow gas to dissociate from the solid gas hydrate structure. Among the various techniques for production of natural gas from gas hydrates, the most economically promising method is considered to be depressurization of an adjacent free gas.

A growing body of evidence suggests that a huge volume of natural gas is stored as gas hydrates in northern Alaska and that production of natural gas from gas hydrates may be technically feasible. However, numerous technical and economic challenges must be resolved before this potential resource can be developed. Many wells have penetrated gas hydrate during oil production operations on the Alaska North Slope. During this time, gas hydrates were known primarily as a drilling hazard. Industry has considered the resource potential of conventional Alaska North Slope gas during industry and government efforts in working toward an Alaska North Slope gas pipeline. Consideration of the resource potential of this conventional Alaska North Slope gas created the industry–government alignment necessary to also consider the resource

potential of the potentially huge (40–100 TCF in-place) unconventional gas hydrate accumulations beneath existing production infrastructure. Two of the wells recently drilled as a part of the industry–government research programs to explore and evaluate gas hydrate production potential on the ANS were the Hot Ice #1 and the Mt. Elbert-01.

### The Hot Ice No. 1

The Hot Ice No. 1 well was drilled on the Anadarko Petroleum leases, southwest of the main Kuparuk development, just outside the Kuparuk River Unit boundary. Drilling began in 2003 and ended in 2004. The well was drilled, cored, and logged to a depth of 2,300 ft (700 m), and ultimately abandoned as no hydrates were found [17]. However, the lessons learned from drilling this well were useful in designing future hydrate exploration wells such as the Mt. Elbert.

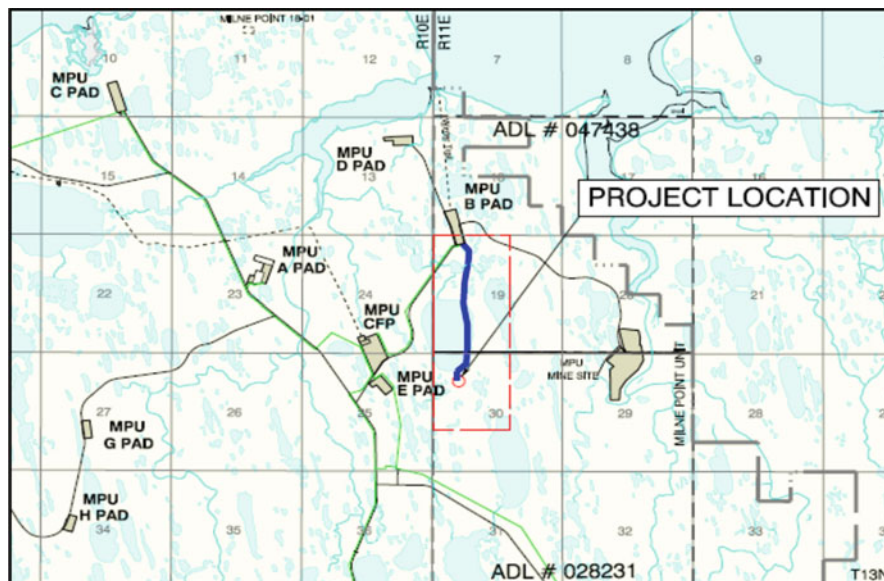
Hot Ice No.1 well also used water-based polymer drilling mud with KCl (salt) for freezing-point depression and a mud chiller for circulating chilled mud. This was problematic, with freeze-ups and mechanical problems impeding progress. This knowledge influenced decisions with regard to coring fluids and mud chilling on the Mt. Elbert-01 Gas Hydrate Stratigraphic Test Well.

### The Mt. Elbert-01

The drilling, coring, and testing of Mt. Elbert-01 Gas Hydrate Stratigraphic Test Well was part of the DOE/NETL-funded “Alaska North Slope Gas Hydrate Reservoir Characterization Project,” and was a cooperative effort between industry, academia, and government.

The Mt. Elbert-01 Hydrate Test well was located on a temporary ice pad in the Milne Point Unit on the ANS, south of the B-Pad and east of the E-Pad, with ice road access from the B-Pad. The location of the Mt. Elbert-01 relative to the Milne Point infrastructure is shown in Fig. 7. The location of the Mt. Elbert-01 Hydrate Well was chosen so as to drill in an area with “known” hydrate accumulations with isolation (stand-off) from the thermal effects of any adjacent wells. This would allow the best chance to find and test undisturbed hydrates.

Based on well log data from nearby offset wells, it was estimated that the Base of the Ice-Bearing Permafrost (BIBPF) was at approximately 1,850 ft TVDss and the Base of the Gas Hydrate Stability Zone (BGHSZ) was at about 2,950 ft TVDss. Coring operations were planned from three potential gas-hydrate-bearing zones (zones D, C, and B) within this depth range. A total of 23 coring runs were made using the wireline



Alaska Gas Hydrate Research and Field Studies. Figure 7

Section Map – Mt. Elbert-01 location relative to Milne Point infrastructure (Courtesy Hanson [18])



retrievable coring method, which prevented the cored formations from being exposed to pipe tripping effects. This was the first application of wireline retrievable coring on the Alaska North Slope. Figure 8 shows a summary of the coring runs. The overall core recovery was 85%. The use of a chilled mineral-oil-based coring fluid and wireline retrievable coring method are believed to be mainly responsible for such high core recovery. The main advantage of a chilled oil-based coring fluid over a chilled water-based coring fluid is that the oil-based fluid would not freeze. Although additives were used, freezing of water-based coring fluids was a major problem in the previous hydrate wells Mallik and Hot Ice #1 [18]. Details of the drilling plan and coring operations for Mt. Elbert-01 are given by Hanson [18].

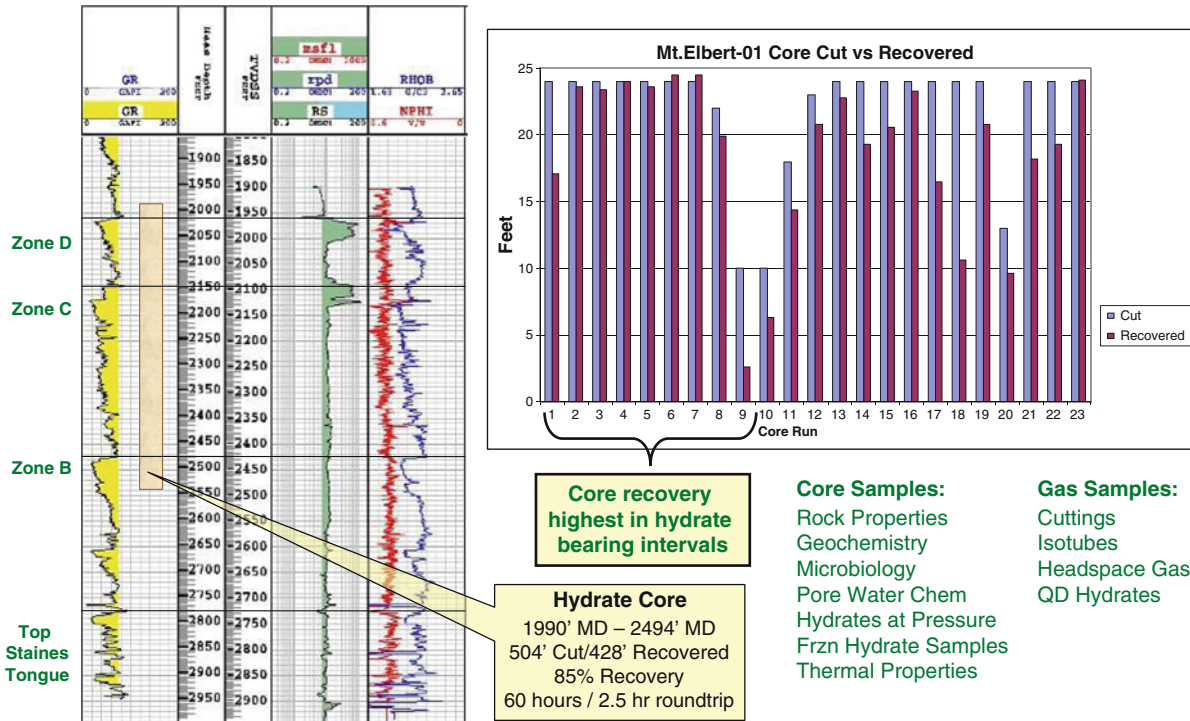
The logging suite for Mt. Elbert-01 consisted of three electrical resistivity tools and an extended MDI testing. Figure 9 shows a sample plot of the well logs, clearly delineating the upper D hydrate zone. Full analysis of Mt. Elbert-01 well logs showed no evidence of significant “contacts” between free gas sands and

hydrate-bearing sediments [18]. Mt. Elbert-01 is the world's first successful application of the MDT™ tool in open-hole in gas hydrates.

## North Slope Borough Methane Hydrate Project Review

Barrow, Alaska, is a remote and the northernmost city in the United States. The city serves both as the economic and administrative center of the North Slope Borough (NSB). The energy demands of the city and nearby localities are met by natural gas supply from the Barrow Gas Fields (BGF) consisting of three different gas pools: the East Barrow (EB) pool, the South Barrow (SB) pool, and the Walakpa (WAL) pool (Fig. 2). The Barrow Gas Fields have been producing natural gas for the last 25 years. It has been estimated that the proven gas reserves in these fields exceed over 250 billion standard cubic foot (BCF) [19, 20]. Based on the current demands for natural gas in the Barrow region and to enhance natural gas production and expand

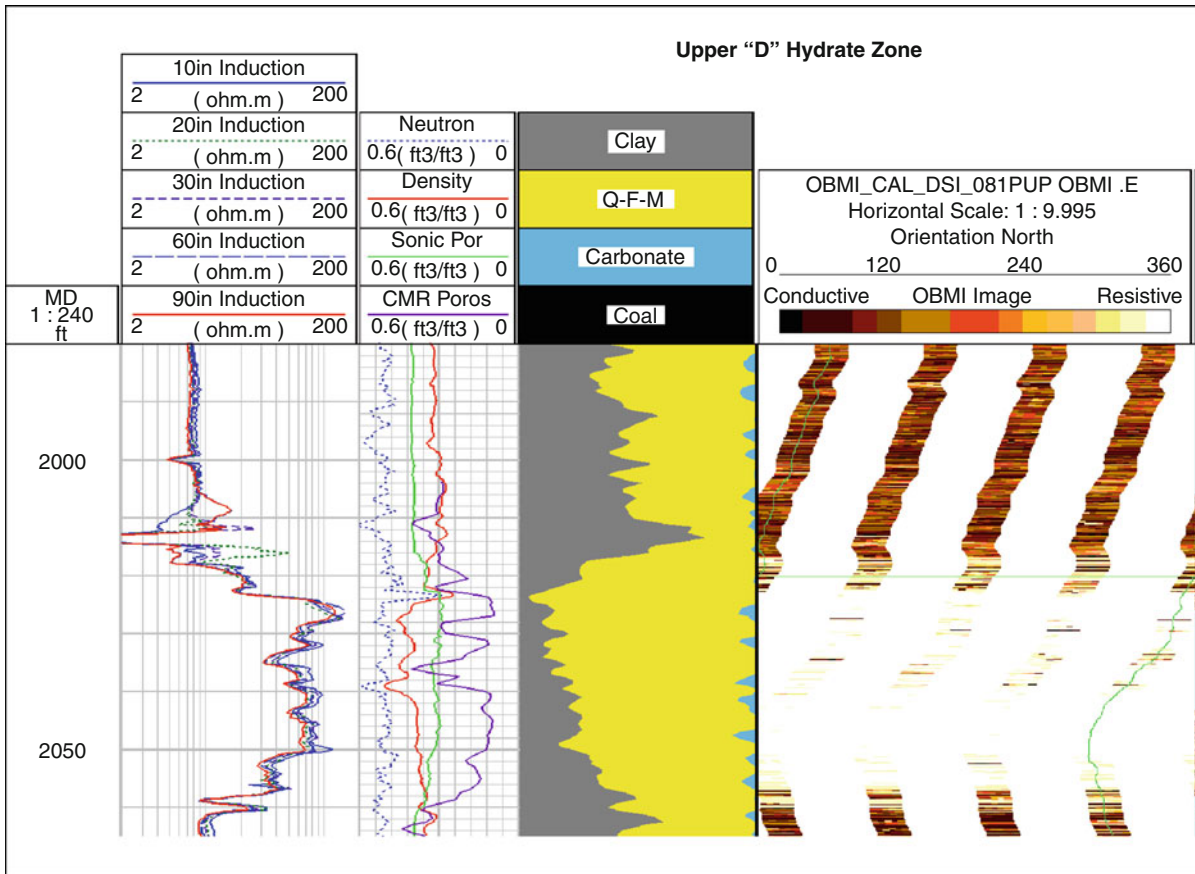
## Summary of Mt. Elbert-01 Core Results



Alaska Gas Hydrate Research and Field Studies. Figure 8

Summary of core recovery by run relative to the log through the hydrate interval (Courtesy Hanson [18])





Alaska Gas Hydrate Research and Field Studies. Figure 9

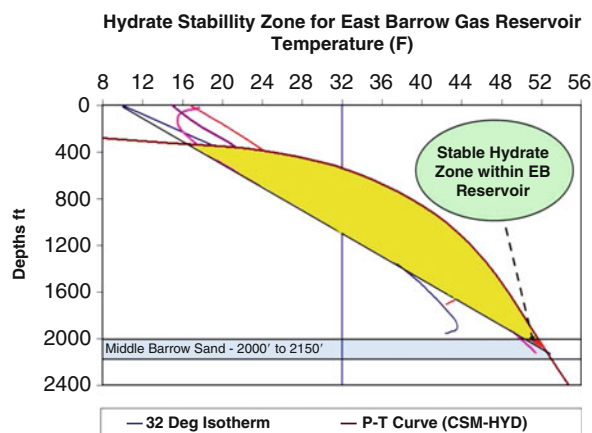
Composite RT Scanner, Neutron, Density, Sonic, CMR and OBMI Log showing the upper “D” – Hydrate interval (Courtesy Hanson [18])

production facilities in order to provide cost effective and assured source of energy to Barrow and nearby areas, the Barrow gas hydrate project was proposed by the NSB to the US Department of Energy (DOE). In 2006, DOE and NSB jointly funded a research program to characterize and quantify methane hydrate resource potential associated with the BGF, with the ultimate goal of producing gas from the hydrate formations associated with individual gas pools of BGF.

#### Gas Hydrate Stability Modeling for BGF Pools

An engineering approach was first adopted to develop methane hydrate stability models for the three gas pools. This was performed on the basis of gas analysis, formation water salinity, and pressure and temperature gradient data available from various sources referenced in [21].

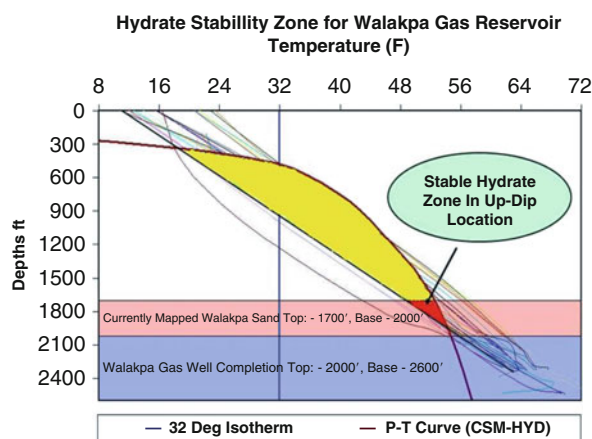
The studied literature indicated a gas composition of practically pure methane (99%) and water salinities in the range of 0–4% w/w. Details of the calculation procedures can be found in [21]. The hydrate stability models for EB and WAL gas pools are provided in Figs. 10 and 11. Results from the EB pool calculations indicate the existence of hydrate stability zones within the pay zone of 2,000–2,150 ft, which suggests that the in situ hydrates are possibly associated with free gas reservoir. Similar to EB, the hydrate stability model for WAL pool also showed promising results in that stable hydrate zones were not found within deeper zones of the WAL sand where existing gas wells were completed within the 2,000–2,600 ft interval but showed stable envelope existing in up-dip locations of the reservoir between 1,700 and 2,000 ft. To confirm these results, detailed geological and geophysical analysis of



### Alaska Gas Hydrate Research and Field Studies.

**Figure 10**

Hydrate stability model for EB pool



### Alaska Gas Hydrate Research and Field Studies.

**Figure 11**

Hydrate stability model for WAL pool (Courtesy Singh [21])

unexplored regions of the pools are required to determine the association of gas hydrates with the prospective gas pools. Singh [21] conducted a material balance study to obtain additional confirmation of the presence of hydrates, especially associated with the EB pools.

### The East Barrow Field Material Balance Study

The objective of the material balance study was to qualitatively understand the behavior of the EB pool under different drive mechanisms, and to confirm if hydrate dissociation was one of them. Stokes and Walsh

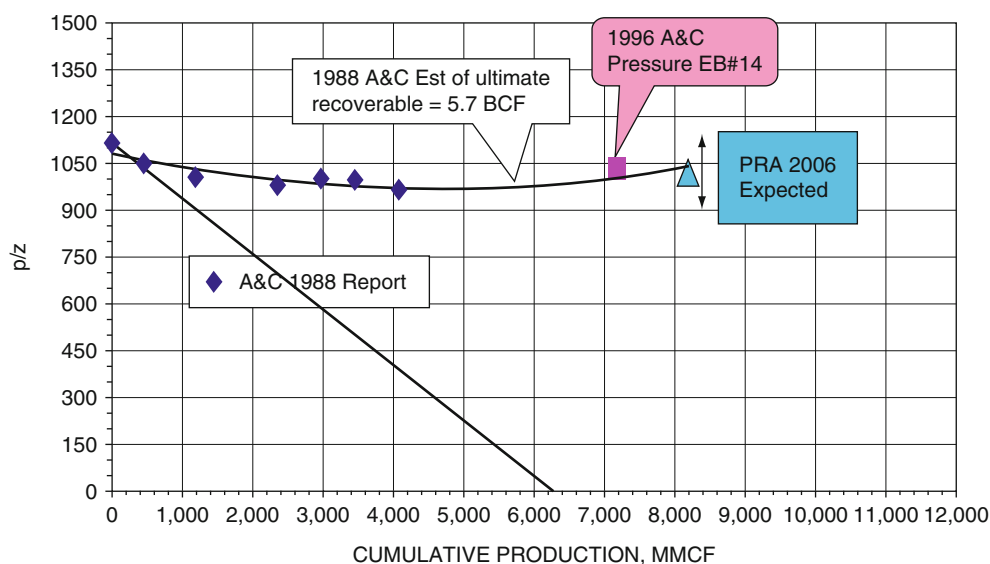
[22] performed pressure buildup tests on two EB wells and later [23] compared the static reservoir pressure data with historical pressure data reported by Allen and Crouch [24]. Figure 12 represents the historical performance plot of the EB gas pool, which shows an initial drop in the static reservoir pressure with gas production, indicating a conventional response of a volumetric gas reservoir producing by gas expansion. However, long-term production data showed gradual pressure stabilization, which indicated the possibility of water influx from a communicating aquifer. But the possibility of a communicating aquifer was ruled out because (1) there was negligible water production ( $\sim 1,600$  bbls up to September 2007), (2) cumulative gas production of  $\sim 8$  BSCF by 2007 was much higher than Allen and Crouch's [24] estimate of  $\sim 6$  BSCF, and (3) none of the EB wells had watered out. Thus, the EB field performance data of Fig. 13 suggested possible contribution from an unknown drive mechanism, which could be gas recharge from dissociating gas hydrates, or water influx from a new aquifer.

In the EB material balance study, the overall performance of a tank type, static reservoir model representing the EB gas pool was compared against the available production data. The following drive mechanisms were studied with the tank model to see which drive mechanism would match actual production data: (1) volumetric gas tank, (2) water drive, (3) dissociating-hydrate drive, and (4) combination of water and hydrate drive. Reservoir and fluid properties and other details are found in Singh [21].

The study showed that the actual production data from the EB field could not be matched by assuming volumetric behavior, water drive, or dissociating-hydrate drive alone. However, a good match of modeled production response and actual production data, shown in Fig. 13 by red dotted circles, was obtained in the scenario where a free gas reservoir was in communication with a weak aquifer, but associated with a thick hydrate layer. Thus, a combination of aquifer support and hydrate dissociation is the most likely drive mechanism in the EB pool.

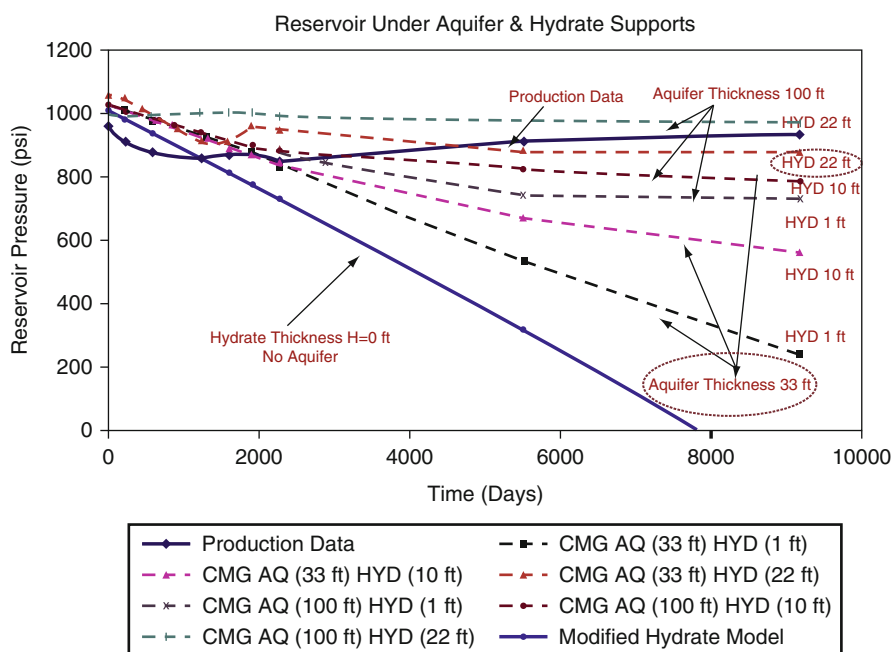
### Field Scale Reservoir Simulation Study

Using the reservoir drive mechanism established through the material balance study on the EB pool,



Alaska Gas Hydrate Research and Field Studies. Figure 12

EB gas pool performance (Stokes and Walsh [23])



Alaska Gas Hydrate Research and Field Studies. Figure 13

Hydrate- and aquifer-supported gas reservoir performance prediction for EB pool (Courtesy Singh [21])

Singh [21] conducted numerical reservoir modeling studies using a full-scale, three-dimensional (3D) dynamic reservoir model. A commercial reservoir simulator (CMG-STARs) was used to perform the

simulation studies. The static reservoir model of Panda and Morahan [25] was imported and built upon by adding reservoir properties and drive mechanisms. Other details can be found in Singh [21].

The simulation study predicted field-wide performance of the EB pool for a period of 30 years and investigated the impacts of well completion and orientation (vertical wells versus horizontal wells). History matching of the simulation results confirmed the previously estimated combination drive as the primary recovery mechanism for this reservoir. However, gas production by expansion dominates over hydrate dissociation mechanism after approximately 15–20 years of production. The simulation study also indicated that infill drilling using horizontal producing wells would offer the ideal combination of lower water production and higher gas recovery.

### Future Directions

Various methods have been studied and have been proved to be successful in the extraction of gas from gas hydrate deposits. These are depressurization, thermal methods, and the use of inhibitors (e.g., methanol or glycols). Gas hydrate dissociation is an endothermic reaction and as the hydrate deposits dissociate, the surrounding temperatures decrease resulting in further hydrate reformation, thereby reducing the amount of gas that can be produced. Thus any method or a combination of methods used for hydrate dissociation has to not only consider the energy required to dissociate hydrates but also has to take in effect the energy that is required to counter the drop in temperature that takes place due to the endothermic nature of hydrate dissociation [26]. Yutaek and SeongPil [27] have shown how kinetic hydrate inhibitors (KHI) can be used to avoid hydrate reformation, the required concentration of these KHI being 0.5–5 wt%.

Recent experiments conducted by Shiang-Tai and Yen-Tien [28] illustrate how replacement of CO<sub>2</sub> with CH<sub>4</sub> in methane hydrates can help in the production of methane gas from methane hydrate deposits. When, liquid carbon dioxide is introduced into the methane hydrate system, rapid replacement of solid methane with carbon dioxide molecules takes place. This substitution of methane with carbon dioxide occurs spontaneously and large amounts of carbon dioxide are observed to penetrate methane hydrate molecules and expel methane especially near the hydrate melting point [28]. It has been shown how steam injection can be used effectively in producing gas from gas

hydrate deposits [29], and the use of hot-brine injection for the production of gas from natural gas hydrates as well as how various factors like the temperature of the injected solvent, the injection rate, or the injection time impact the production of gas from such deposits has also been shown [30]. It is also proposed that a combination of thermal and depressurization technique can be used to produce from gas hydrates that will help in improving the economic efficiency, paving the way for the development of NGH reservoirs [30].

Weitemeyer and Constable [7] have shown how controlled source electromagnetics can be used to map gas hydrates. Further advancement of this technique has led to the creation of a fixed offset towed sensor that was able to image gas hydrates at MC118 with successful imaging of hydrates in the other three sites in the Gulf of Mexico [7]. Dai [31] has put forward a method to locate and quantify gas hydrate deposit using long-offset reflection seismic information, and since this method is not based on the AVO inversion method it stays away from the uncertainties that are a part of AVO-based seismic studies.

### Conclusions

Significant improvements in the methods for drilling and evaluating potential gas hydrate zones in ANS have taken place in recent years. The drilling and formation evaluation of the Mt. Elbert-01 stratigraphic test well has successfully field-tested the modern geophysical prospecting methods used and has enabled the selection of target zones and field parameters for potential future production testing. The chilled, 100% mineral-oil-based coring fluid provided a significant improvement in wellbore quality as compared to the freezing-point-depressed water-based fluids used in the previous gas hydrate exploration projects. Core recovery, core quality, and electrical well log quality were significantly improved by using oil-based coring fluid, and operational downtime due to freeze-ups was eliminated.

The study of a producing gas field in the ANS has demonstrated the feasibility of producing natural gas from gas hydrates by depressurization to some extent. Diagnostic methodologies were developed to examine the production data from the Barrow gas fields to test the two fields for possible hydrate accumulations and to understand the gas recovery mechanisms. Engineering

approach was followed to construct hydrate stability models and material balance procedures for identifying the reservoir drive mechanisms. Based on these results, state-of-the-art dynamic reservoir simulation models were developed using an advanced commercial reservoir simulator to study hydrate reservoir development. Reservoir modeling confirmed the key findings of previous studies and projected the future performance of the reservoir under different scenarios. Finally, hydrate resource potential of EB and WAL pools appears to be promising, and both pools have considerable hydrate potential that can supply large volumes of gas for future years.

### Disclaimer

This entry was prepared as an account of work sponsored by an agency of the US government. Neither the US government nor an agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trademark, manufacture, or otherwise do not necessarily constitute or imply its endorsement, recommendation, or favoring by the US government or any agency thereof. The views and opinions of the author expressed herein do not necessarily state or reflect those of the US government or any agency thereof.

This entry is also based on research conducted by two petroleum engineering graduate students (Paul Hanson and Praveen Kumar) from the University of Alaska Fairbanks for their M.S. degree in petroleum engineering. The results of the two US Department of Energy-funded gas hydrate projects in Alaska (BP Alaska Gas Hydrate Project and the North Slope Borough Gas Hydrate Project) are also included. Several other companies, organizations associated, and individuals associated with these two projects have contributed to the work reported here.

### Acknowledgment

The authors are grateful to the US Department of Energy (USDOE), BP Exploration (Alaska) Inc., North Slope Borough (NSB), Petrotechnical Resources of Alaska

(PRA), and Petroleum Development Laboratory (PDL) at UAF for their financial support in accomplishing this work. The authors would also like to thank Dr. Tim Collet (USGS), Mr. Robert Hunter (ASRC Energy Services), Mr. Tom Walsh (Petrotechnical Resources Alaska, Inc.), and graduate students Praveen Singh, Andrew Johnson, Paul Hanson, and Vivek Peraser for their contribution, support, and help when needed.

### Bibliography

1. Folger P (2010) Gas hydrates: resource and hazard. CRS report for congress, May 2010, CRS, Washington, DC
2. WEC (2007) Survey of Energy Resources, 2007. World Energy Council, London
3. Islam MR (1991) A new recovery technique for gas production from Alaskan Gas Hydrates. SPE 22924, Dallas, TX, USA
4. Garg R, Ogra K, Choudhary A, Menezes R. Chemical recovery of gas hydrates using fluorine gas and microwave technology. SPE 113556, TX, USA
5. Peter E (1993) Clathrate Hydrates. Ind Eng Chem Res 32:1251–1274
6. Collett TS (1992) Geologic comparison of the Prudhoe Bay-Kuparuk River (U.S.A.) and Messoyakha (U.S.S.R.) gas hydrate accumulations. SPE 24469, Texas, USA. pp 1–20
7. Weitemeyer K, Constable S (2011) Mapping gas hydrates with marine controlled source electromagnetics. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
8. George MJ, Timothy CS, Boswell R, Kurihara M, Matthew RT, Koh C, Dendy SE (2008) Toward production from gas hydrates: current status, assessment of resources, and simulation based evaluation of technology and potential. SPE 114163, Texas, USA
9. Sloan ED, Koh C (2008) Clathrate hydrates of natural gases, 3rd edn. Taylor and Francis, Boca Raton, FL
10. Johnson Arthur H (2011) Global resource potential of gas hydrate – a new calculation. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
11. Boswell R, Collett TS (2006) The gas hydrates resource pyramid. In: Fire in the ice, methane hydrate newsletter. US Department of Energy, Office of Fossil Energy, National Energy Technology Laboratory, Fall Issue, pp 5–7
12. Ruppel C (2011) Methane hydrates and the future of natural gas. In: MITEI Natural gas Report, MIT, MA (Supplementary paper on methane hydrates)
13. Nagakubo S, Arata N, Yabe I, Kobayashi H, Yamamoto K (2011) Environmental impact assessment study on Japan's methane hydrate R&D program. In: NETL, Methane Hydrate Newsletter, January 2011
14. Collett TS, Johnson AH, Knapp CC, Boswell R (2009) Natural gas hydrates. A review. AAPG Memoir 89:146–219
15. Paull CK, Matsumoto R (2000) Leg 164 Overview. In: Proceedings of the ocean drilling program, Scientific Results, vol 164. Ocean Drilling Program, TX, USA



16. Pecher IA, GHR Working Group (2011) Gas hydrates in New Zealand – a large resource for a small country? In: NETL, Methane Hydrate Newsletter, January 2011
17. Kadaster AG, Millheim KK, Thomson TW (2005) The planning and drilling of hot ice #1 – gas hydrate exploration well in the Alaskan Arctic. In: SPE 92764-MS presented at the SPE/IADC drilling conference, 23–25 February 2005, Amsterdam, Netherlands, pp 1–15
18. Hanson P (2007) The Mt. Elbert-01 Stratigraphic test well – lessons from drilling, coring and evaluating gas hydrates on the Alaska North Slope, MS project report, University of Alaska Fairbanks, Fairbanks, December 2007
19. Gruy HJ (1978) Reservoir engineering and geologic study of the East Barrow Field, National Petroleum Reserve in Alaska. Technical Report submitted to Husky Oil NPR Operations, Inc., Anchorage, Alaska, pp 1–62
20. Darkwah SA, Allen WW (1996) Engineering study of the South, East and Walakpa Fields, North Slope Borough, Alaska. Technical Report submitted to the North Slope Borough, Barrow, Alaska, pp 1–32
21. Singh PK (2008) An engineering study to investigate methane hydrate resource potential associated with barrow gas fields, Alaska. MS Thesis, University of Alaska Fairbanks, Fairbanks
22. Stokes PJ, Walsh TP (2006) South and East Barrow gas fields well improvements study. Technical Report submitted to the North Slope Borough, Barrow, Alaska, pp 1–70
23. Stokes PJ and Walsh TP (2007) South and East Barrow gas fields reserves study. Technical Report submitted to the North Slope Borough, Barrow, Alaska, pp 1–36
24. Allen WW, Crouch WJ (1988) Engineering study of South and East Barrow fields, North Slope Alaska, Alaska. Technical report prepared for North Slope Borough Gas Development Project, Barrow, Alaska, pp 1–56
25. Panda MN, Morahan GT (2008) An integrated reservoir model description for East Barrow and Walakpa gas fields. Topical Report prepared under DoE Project No. DE-FC26-06NT42962 submitted to the USDoE, Morgantown, WV, [www.netl.doe.gov](http://www.netl.doe.gov), pp 1–44
26. Sandilya P, Ganguly S (2011) Conceptual design of an experimental facility for extracting natural gas from natural gas hydrate. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
27. Yutaek S, Seong-Pil K (2011) Dependence of drawdown pressure on the hydrate reformation during methane hydrate production and its inhibition with kinetic hydrate inhibitors. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
28. Shiang-Tai L, Yen-Tien T (2011) Methane hydrate growth and its recovery with carbon dioxide via molecular dynamic simulations. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
29. Lee Y, Shin C, Baek Y, Kim Y, Jang S, Lee J (2011) A study on the dissociation behavior of gas hydrate using the steam stimulation. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
30. Shuxia L, Yueming C, Yongmao H (2011) Experimental study of influence factors of hot-brine stimulation for dissociation of NGH in porous medium. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
31. Dai J (2011) A method to locate and quantify gas hydrate using long-offset seismic. In: Proceedings of the 7th international conference on gas hydrates (ICGH 2011), Edinburgh, UK
32. Collett TS (1993) Natural gas hydrates of the Prudhoe Bay and Kuparuk River area, North Slope, Alaska. AAPG Bull 77(5): 793–812

---

## Algae, a New Biomass Resource

CINZIA FORMIGHIERI, ROBERTO BASSI

Dipartimento di Biotecnologie, Università di Verona, Verona, Italy

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Advantages of Algae as Biomass Producers  
 Present Algal Productivity in the Laboratory Versus Large Scale  
 Domestication  
 Light Use Efficiency  
 Non-Photochemical Energy Quenching at the Molecular Level  
 Interconversion and Storage of Photosynthetic Metabolic Products  
 Screening for State Transition as an Indirect Mean to Select Strains with Altered Redox Metabolism  
 Accumulation of Biomass as Neutral Lipids  
 Planning an Algal Refinery  
 Large-Scale Systems  
 Future Perspectives and Technological Developments  
 Bibliography

### Glossary

**Algae** Oxygenic photosynthetic organisms, prokaryotic or eukaryotic, with organization ranging from unicellular to multicellular. Algae never have true stems, roots, and leaves, thus leading to their classification as “lower” plants.

**Biofuel** Renewable energy-rich compound derived from living organisms or from their metabolic by-products.

**Biomass** Organic raw material, stored as a result of the metabolism of a living organism, which can be used as a resource for energy and biofuels.

**Photoautotrophy** The ability of a living organism to use carbon dioxide as carbon source for biomass and light as source of energy.

**Photosynthesis** The overall process that converts light energy into chemical energy, finally used to fix inorganic carbon dioxide into organic compounds needed for growth. In oxygenic photosynthesis, light-induced redox reactions occurring in the photosynthetic electron transport chain are coupled to the extraction of electrons from water.

**Photosystem** A multipigment-protein complex composed of a light-harvesting antenna moiety energetically connected to a reaction center where the excitation energy is used for charge separation and production of a reduced product.

**Light saturation constant** The intensity of light at which photosynthetic oxygen evolution and specific biomass growth rate are half the maximum level. The energy absorbed in excess with respect to the photosynthesis saturation is dissipated as fluorescence or heat and not used for photochemistry.

**Photoinhibition** The light-induced inactivation of photosynthesis occurring when photooxidative damage of the photosynthetic machinery (particularly photosystem II) overcomes the capacity for repair.

**Genetic improvement** All procedures, including phenotypic selection, conventional breeding, mutagenesis, and genetic engineering, aimed at indirectly or directly influencing the genetic background of a wild strain, which was evolved following rules of natural selection. Genetic improvement is intended at improving existing characteristics or at introducing new traits to fit applications.

**Genetic engineering** All the techniques of recombinant DNA to directly manipulate genotypes.

## Definition of the Subject

Algae are oxygenic photoautotrophs, offering a very high level of biodiversity and thus suitable for different

practical applications. Today, they are mainly cultivated for human/animal food or to extract high-value chemicals and pharmaceuticals. However, their exploitation could be extended. Algae are attractive as high yield biomass producers, because of the short life cycle, the ability to grow up to very high cell densities, and the easy large-scale cultivation that does not compete with other demands such as those of conventional crops agriculture. Algae can be a resource of renewable, sustainable biofuels. In addition, they can be transformed into “cell factories” to produce recombinant proteins of interest for pharmaceutical companies.

## Introduction

Algae are described as “lower” plants that never have true stems, roots, and leaves, and grow photoautotrophically by performing oxygenic photosynthesis [1]. They are mostly eukaryotic, although prokaryotic cyanobacteria are included in algae. More than 200,000 species are estimated and form a highly heterogeneous group, spread in all aquatic ecosystems but also in other habitats such as soil. A broad spectrum of phenotypes and specialized adaptation abilities exists within this group while members colonize diverse ecological habitats, from freshwater to marine and hyper-saline, at different temperatures, pH, and nutrient availabilities [1, 2]. Algae are classified as follows: cyanobacteria (Cyanophyceae), green algae (Chlorophyceae), diatoms (Bacillariophyceae), yellow-green algae (Xanthophyceae), golden algae (Chrysophyceae), red algae (Rhodophyceae), brown algae (Phaeophyceae), dinoflagellates (Dinophyceae), and “pico-plankton” (Prasinophyceae and Eustigmatophyceae). Algae are accountable for the net primary production of ~50% of the total organic carbon produced on earth each year [3], a reason for their massive biological importance. Long before the advent of biotechnology, algae were used as a human food source for centuries by indigenous populations [4, 5]. However, aside from some attempts to cultivate algae and to extract from them valuable products back to the seventeenth century [6], commercial large-scale production of algae began few decades ago [4, 7]. Currently, algae are mainly cultivated as human/animal food source and used in aquaculture, in agriculture, as fertilizers, and in order to produce high-value

chemicals, pharmaceuticals, and cosmetics, such as polyunsaturated  $\omega 3$  fatty acids, proteins, biopolymers, and polysaccharides as agar, carrageenan, alginates, pigments, vitamins, and antioxidants [1, 4].

Recently, a new interest in algal biomass as renewable energy source is emerging. About 80% of world energy demands are met by the combustion of fossil fuels: coal, oil, and natural gas. However, fossil fuel reserves are limited and their exhaustion is estimated, assuming a stable consumption, to occur within few decades while their combustion has high impact on environment, due to releases of large amounts of carbon dioxide and other pollutants. This is considered a major problem for human society. Its becoming increasingly clear that finding alternative, renewable energy sources to sustain our lifestyle is striking and urgent. Such widespread and environmentally sustainable energy source will allow single countries to be more independent from the escalating prize of crude oil. The public conscience that search for energy sources alternative to fossil fuels is fundamental for future prospects is generally increasing and research is undertaken in many countries. Advances have been made recently on the exploitation of renewable energies such as sun and wind, which can be considered as “clean” and non exhaustible. Fuels derived from living organisms or from their metabolic by-product are defined as “biofuels.” Photosynthetic organisms convert carbon dioxide into biomass, a form of stored chemical energy that can be used in substitution for fossil fuels. In some species of algae, 30% up to 80% of the biomass is accumulated as oil that can be extracted and converted to biodiesel [2, 8–11]. Some algae, mainly green, can photobiologically generate hydrogen [12–15], which is released from the culture with a purity up to 98% [16]. Spent algal biomass can be further valorized to produce biomethane from anaerobic digestion of the biomass itself [1, 5] and high-value ingredients for food and pharmaceuticals can be produced along with biofuels in order to reach cost-effectiveness and economic competitiveness over conventional fuels.

### Advantages of Algae as Biomass Producers

The use of algae for fuel production has notable advantages. First of all, algae are photoautotrophic

organisms, thus able to produce biomass from solar energy, water, and carbon dioxide, all renewable and cheap components. Carbon dioxide removal, in addition, would contribute to reduce atmospheric levels of this pollutant which, instead, is stoichiometrically released in the atmosphere by conventional fuels. The use of fossil fuels currently releases about 6 billion tons of carbon dioxide per year, only partially compensated by re-assimilation by photosynthetic organisms in forests (about 1 billion ton year<sup>-1</sup>) and oceans (2 billion tons year<sup>-1</sup>) [17] leading to a 3-billion-ton year<sup>-1</sup> increase in the global level of carbon dioxide in the atmosphere. This has been widely suggested to cause the so-called greenhouse effect, leading to world global warming [18]. To prevent undesired climate changes, carbon dioxide emissions need to be reduced in the near future [19]. Biofuels from algae are attractive because their combustion would not result in net carbon dioxide emissions since biofuels will only release the amount of carbon previously absorbed during growth of algae.

Advantages of algae over crop plants are many:

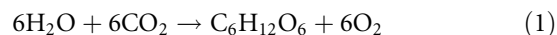
1. Photosynthesis increases with carbon dioxide concentration until saturation. Algae can grow with up to 18% of carbon dioxide, being more efficient carbon dioxide assimilators than land plants [20] and suitable for mitigation of carbon dioxide emissions when coupled to an industrial activity. *Chlorella vulgaris* was shown to respond well to elevated carbon dioxide levels (over 1,850 ppm CO<sub>2</sub>) and to effectively remove up to 74% of the carbon dioxide in the airstream to ambient levels (330 ppm CO<sub>2</sub>) with a 63.9 g m<sup>-3</sup> h<sup>-1</sup> bulk removal in the experimental photobioreactor of 2 L [21]. The carbon dioxide fixation rate by algal biomass is estimated ten times as large as that of the temperate forest [22].
2. Cultivation of algae does not compete with other demands such as food production because algae can grow in wastelands rather than in arable lands. Using the entire US soybean crop for biodiesel would replace only 10% of conventional diesel consumed and total world plant oil production would only satisfy 80% of US demand [23]. Exploiting algae rather than crops thus represents an essential, sustainable alternative to fulfill the energetic demands without affecting food supply.

3. Algae, differently from crops, do not require pesticides, which avoids contamination of water and soils and also fertilization is limited to the culture vessel without dispersion of nutrients in the environment and consequent eutrophication of water bodies.
4. Algae can grow in seawater, brackish water, or wastewater making a substantial saving of freshwater that in contrast is required by conventional crops in high amounts. Extensive cultivation of algae would therefore be more environmentally sustainable than extensive cultivation of conventional crops.
5. Algae can utilize nutrients from a variety of wastewater sources (agricultural runoff, and industrial and municipal wastewaters), providing additional benefit of bioremediation.
6. The life cycle of algae is more rapid (hours to days) as compared to land plants (months to years), allowing for a faster turnover and a higher biomass yield. In more detail, crop plant life cycle has a long initial development phase using energy stored in seeds for building the organism, followed by a relatively short phase of full photosynthetic activity and then senescence, limiting the period of efficient light harvesting and utilization to a fraction of the year. In contrast, algae not only have a short life cycle but they remain productive throughout the year, allowing for a continuous biomass production when light is available and temperature allows.
7. Many algal species are competent for using organic carbon as energy source. This ability opens the possibility to allow for biomass production at night by supplying low-priced carbon substrates.
8. Algal biomass is fully photosynthetically active. In contrast, photosynthesis is localized exclusively in leaves of vascular plants, a fraction of the total plant body, while the rest is an energy sink, thus a disadvantage in terms of biomass productivity.

### Present Algal Productivity in the Laboratory Versus Large Scale

Algae can convert solar energy into chemical energy through the process of photosynthesis. The whole

process of oxygenic photosynthesis can be summarized by Eq. 1.



Light energy is used to extract electrons from water ( $\text{H}_2\text{O}$ ), thus generating oxygen ( $\text{O}_2$ ). These electrons are transported through a linear electron transfer chain and finally reduce  $\text{NADP}^+$  to NADPH. Photosynthetic electron transport is coupled to the generation of a transmembrane electrochemical gradient, whose stored energy is used to synthesize ATP. NADPH and ATP are then used to produce glyceraldehyde-3-phosphate from  $\text{CO}_2$  in a metabolic pathway that is called the Calvin–Benson cycle [24]. Photosynthesis enables the cell to convert inorganic  $\text{CO}_2$  into organic carbons and finally to accumulate biomass.

A theoretical estimation of algal biomass yield can be assessed considering the efficiency of photosynthesis. A minimum of 8 photons of light energy are absorbed per each oxygen molecule evolved, the actual average measurement is 9.5 photons per oxygen [25]. Per oxygen evolved, four electrons are channeled into the linear photosynthetic electron transport chain and in the process three molecules of ATP and two of NADPH are produced. In order to convert three carbon dioxide molecules into glyceraldehyde-3-phosphate, nine ATP and six NADPH are required. The ratio is 9.5 mol photons for the conversion of 1 mol  $\text{CO}_2$  into biomass. The average insolation, that is the solar radiation energy on a surface area in a given time, is between 3 and 5  $\text{kWh m}^{-2} \text{day}^{-1}$  (full solar spectrum) at temperate regions. However, only about 40% of solar radiation is photosynthetically active (PAR), because only photons with wavelengths between 400 and 700 nm (visible spectrum) can be absorbed by photosynthetic pigments, since these wavelengths carry an energy equal to the change in the energetic level between the ground state and the excited one.

A photon has an energy that is directly proportional to its frequency and inversely proportional to its wavelength, following Eq. 2.

$$E = h\nu = c/\lambda \quad (2)$$

( $h$ , Planck constant,  $6.626 \times 10^{-34}$  Js;  $c$ , light speed,  $3 \times 10^8$   $\text{m s}^{-1}$ ;  $\nu$ , frequency;  $\lambda$ , wavelength)

For example, 1 mol photons (1 Einstein) at 440 nm has energy of 272 kJ, while 1 mol photons at 670 nm has

energy of 178 kJ. Out of the full solar radiation of  $5 \text{ kWh m}^{-2} \text{ day}^{-1}$ , the photosynthetically active radiation (PAR) is about  $35 \text{ mol photons m}^{-2} \text{ day}^{-1}$ . Considering the energy requirement of 9.5 mol photons per 1 mol  $\text{CO}_2$ , with such average available light radiation the cell could assimilate  $3.68 \text{ mol CO}_2 \text{ m}^{-2} \text{ day}^{-1}$ . Since biomass composition can be approximated to the formula  $\text{CH}_2\text{O}$ , the previous data would translate in the synthesis of about  $110 \text{ g biomass m}^{-2} \text{ day}^{-1}$ . However, accumulation of organic carbons as biomass is lower, due to respiration and other metabolic activities and energy losses are accounted to 30%. The resulting expected maximum biomass productivity would be of  $77 \text{ g biomass m}^{-2} \text{ day}^{-1}$  ( $280 \text{ t ha}^{-1} \text{ year}^{-1}$ ), corresponding to a solar-to-biomass conversion efficiency of 8–10% [25]. This value is about 25 times higher than estimations for vascular plants, possibly due to algal biomass being fully photosynthetically active. Such estimation assumes that all available light energy is absorbed and utilized for photosynthesis. However, real algal biomass productivity achieved so far in the laboratory or small-scale systems does not exceed the  $73\text{--}146 \text{ t dry weight ha}^{-1} \text{ year}^{-1}$  ( $20\text{--}40 \text{ g dry weight m}^{-2} \text{ day}^{-1}$ ) and 3% of solar-to-biomass conversion efficiency [25] in the best cases. Currently, commercial exploitation of algae mainly utilizes open ponds. They are usually raceway cultivators driven by paddle wheels or unstirred, operating at water depths of 15–20 cm, biomass concentration can be up to  $1,000 \text{ mg L}^{-1}$  and productivity between 60 and  $100 \text{ mg L}^{-1} \text{ day}^{-1}$  ( $10\text{--}25 \text{ g m}^{-2} \text{ day}^{-1}$ ) [26]. Commercial rates of production of *Chlorella* have been reported to be  $60\text{--}75 \text{ t dry weight ha}^{-1} \text{ year}^{-1}$  ( $17\text{--}20 \text{ g dry weight m}^{-2} \text{ day}^{-1}$ ) [27]. Another example is *Dunaliella salina*, which is now being cultivated on a commercial scale and in pilot-scale projects to extract  $\beta$ -carotene and glycerol. Over short periods and in small-scale, productivity records  $60 \text{ g dry weight m}^{-2} \text{ day}^{-1}$ . However, commercial production requires larger culture systems and is presently carried out in outdoor ponds (approximately 20 cm deep for 5 ha of surface, total ponds area of 50 ha), because of the still expensive constructing of closed photobioreactors. In these large-scale systems, the maximum productivity achievable is at present  $30\text{--}40 \text{ g dry weight m}^{-2} \text{ day}^{-1}$  but more regularly is below  $30 \text{ g dry weight m}^{-2} \text{ day}^{-1}$  [28]. Outdoor pond cultures of *Cyclotella* report a biomass productivity of

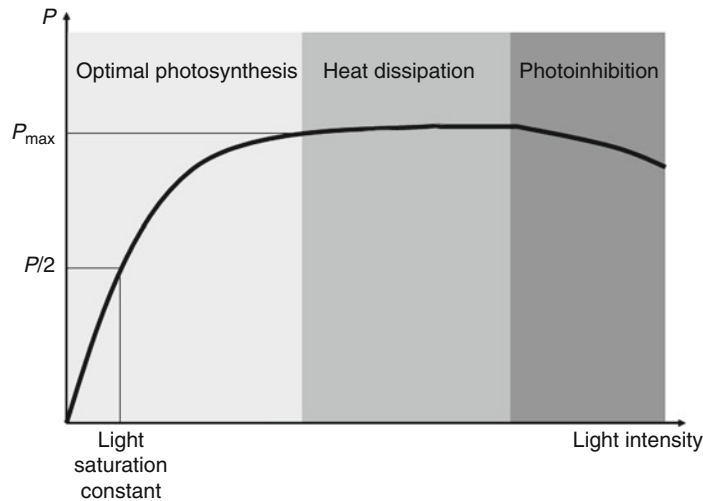
$12 \text{ g m}^{-2} \text{ day}^{-1}$  [29]. With the present knowledge and with the available algal strains, real productivity is far below theoretical estimations, especially in large-scale and over long-lasting periods, and a major future goal would be to get rid of or reduce this gap.

As an alternative to open ponds, few relatively large-scale closed photobioreactors have been developed (refer to “Large-Scale Systems” section for a comparison between open ponds and photobioreactors in terms of advantages), displaying a productivity of  $2.7 \text{ g L}^{-1} \text{ day}^{-1}$  in a small undular row tubular photobioreactor of 11 L,  $1.9 \text{ g L}^{-1} \text{ day}^{-1}$  in a airlift tubular photobioreactor of 200 L and  $0.05 \text{ g L}^{-1} \text{ day}^{-1}$  in a parallel tubular photobioreactor of 25,000 L [30]. Constructing of photobioreactors is still at the beginning and better productivities (up to ten times higher  $\text{g L}^{-1} \text{ day}^{-1}$ ) as compared to open ponds are obtained with pilot systems. It is likely that such interesting results would also be achieved at larger scale, by improving photobioreactor design and mechanics.

One major reason for the real lower biomass productivity with respect to the theoretical calculation is that photosynthesis light reactions have to fit downstream biochemical processes while excess absorbed energy is dissipated as heat. Photosynthesis displays a light saturation curve [25] (Fig. 1).

At low light intensities, light is the limiting factor for the photosynthesis rate, measured as photosynthetic oxygen evolution. At increasing light intensities, the limiting factor becomes carbon dioxide fixation. When light irradiance overcomes the rate of the downhill biochemical processes, excess absorbed energy is dissipated and photosynthetic oxygen evolution reaches a plateau. If light further increases, beyond the capacity of photoprotective mechanisms, photoinhibition leads to a decrease in photosynthesis rate. Therefore, all the absorbed light energy in excess with respect to the saturation level of photosynthetic electron transport does not contribute to biomass accumulation, but it is instead wasted as heat and/or leads to the formation of reactive oxygen species (ROS) that ultimately inhibit photosynthesis. Light saturation constant is defined as the intensity of light at which the specific biomass growth rate is half its maximum value [9]. For example, light saturation constants of microalgae *Phaeodactylum tricornutum* and *Porphyridium cruentum* are, respectively,  $185 \mu\text{mol photons m}^{-2} \text{ s}^{-1}$  [31] and  $200 \mu\text{mol}$





**Algae, a New Biomass Resource. Figure 1**

Light saturation curve. Photosynthetic oxygen evolution ( $P$ ) increases with light intensity linearly until saturation.  $P_{max}$  is the maximum photosynthetic rate.  $P/2$  is the photosynthetic oxygen evolution rate at half the maximum level. The light intensity at  $P/2$  is defined as the light saturation constant. Upon saturation of photosynthesis, when downstream biochemical processes are limiting, excess absorbed energy is dissipated as heat and if light further increases beyond the capacity of photoprotective mechanisms, photosynthetic rate decreases because of photosystem II photoinhibition

photons  $\text{m}^{-2} \text{s}^{-1}$  [32], much lower than the maximum outdoor sunlight level that occurs at midday in equatorial regions, that is about 2,000  $\mu\text{mol photons m}^{-2} \text{s}^{-1}$  [9]. In particular, considering a saturating light intensity of 400  $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ , photosynthesis would saturate at about 7 a.m. and remain saturated until 5 p.m. [25]. An average of 60% up to more than 80% of absorbed irradiance would be wasted and not converted into chemical energy during the course of a sunny day [25, 33]. These energy dissipation events have an important photoprotective role in the natural environment but reduce the potential growth rate in biomass culture conditions.

## Domestication

Crop species currently widely cultivated are “domesticated” strains that evolved from wild ancestors [34]. A domesticated crop has been genetically altered and made into a resource for man through cycles of phenotypic selection, breeding, and mutagenesis. In this way, man can divert evolution of plants to fit agronomical, nutritional, and industrial applications. A fully domesticated crop cannot survive without the help of human mankind; meanwhile, farmers would

not work with wild species because farming with wild-type (ancestral) genotypes would not be economically sustainable. Domestication began in Middle East around 10,500 years ago when humans applied selection to cereals and legumes, modifying morphological and nutritional traits. During history, all crop plants were domesticated in a process that took between decades and several centuries depending on the number of genes involved. For example, modern corn was domesticated from the wild *Teosinte* to improve productivity and the ability to grow at high density in cornfields. High plant densities in fields are not dissimilar to the high algal cell densities that are reached up in a photobioreactor. In the case of tomato, several mutations in genes involved, for example, in cell division, carotenoid biosynthesis, anthocyanin biosynthesis, and ethylene receptor altered fruit color, weight, shape, and ripening of the wild ancestor to finally obtain the now commercially cultivated tomato [35]. Indirect conventional approaches are now being supported by recombinant DNA techniques to directly manipulate genotypes and speed up the process.

The natural biodiversity of algae offers a wide spectrum of phenotypes and specialized adaptation abilities that can be exploited for commercial applications.

Relatively few algal strains have been examined to date, among all the species available in nature. Strains with very interesting features as accumulators of biomass, biofuels, or other high value products could be still unexplored and search for new strains is for sure a valuable strategy to pursue. However, similarly to crops, biofuel production with wild-type algal species collected from the environment, often proposed as a “green way” to energy supply, is unlikely to be economically sustainable, because wild-type algal strains evolved to better adapt to their natural habitats and some of their characteristics are far from optimal for growth in mass culture conditions. A domestication process needs to be applied to algal species preliminary to industrial applications are attempted. In particular, selection of strains with the desired properties combined with input of new alleles through mutagenesis and genetic engineering would help to generate strains with improved biomass yield, oil content, and fuel properties. For example, insertional mutagenesis can be used to generate random insertional libraries to be screened for selected phenotypes.

Algal biotechnology primarily utilizes unicellular species that can be propagated in the laboratory. For this reason, a great attention is focused on unicellular microalgae, more suitable for genetic manipulation and prospects of biofuels production [1]. Algae have several advantages over higher plants in genetic improvement:

1. A short life cycle enables to genetically manipulate them faster than crop plants.
2. Because of the usual absence of cell differentiation and haploid nature of most vegetative stages, microalgae allow faster phenotypic selection.
3. Microalgae are small and can be analyzed in large numbers in a Petri dish.
4. It is conceivable that substantial improvements in algae cultivation could be achieved in the future. In contrast, prospects of improvement in higher plants are less favorable, because current cultivated crops have already been genetically improved with respect to the wild ancestors and cultivation techniques are already well optimized. Algae have been less explored and exploited in the past and current biomass productivity is likely to be improved.
5. Algae are suitable for growth in photobioreactors, offering a confined environment for genetically modified organisms.

In the following sections, examples of targets for domestication are described.

### Light Use Efficiency

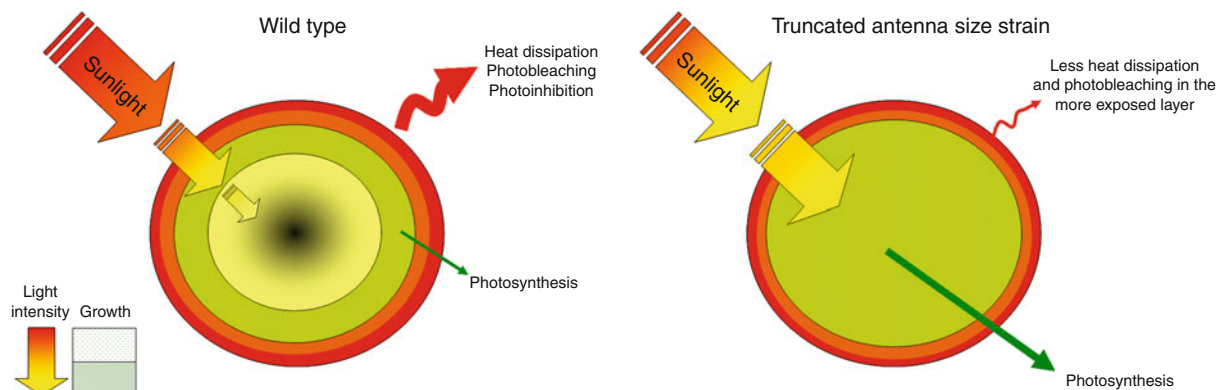
A main process affecting solar-to-biomass conversion efficiency is light harvesting. The photosynthetic apparatus comprises two photosystems (photosystem II and photosystem I) operating in series. Each photosystem is composed of an essential core complex, decorated with an antenna system of variable size with functions in light harvesting and photoprotection. Chlorophyll molecules and other accessory pigments bound to photosystems act cooperatively in the absorption of incoming solar radiation. The ability to assemble large arrays of light-harvesting complexes has been positively selected during evolution since it represents a survival strategy and a competitive advantage in the wild, where light could be limiting [36]. Meanwhile, excess light conditions are mostly avoided by algal cells through changing their location in the water column. Up to 600 chlorophyll molecules can be found in association with photosystem II and photosystem I [25]. Light-harvesting complexes are however not a rigid apparatus and both short-term and long-term mechanisms to adjust the light-harvesting capacity to changing light conditions are present in photosynthetic organisms. These include heat dissipation of excess absorbed energy (non-photochemical quenching, NPQ) [37, 38] in the short term and regulation of photosynthetic gene expression during long-term acclimation [39–41]. Growth conditions in large-scale mass culture are very different with respect to those found in the natural environment. In photobioreactors, cellular concentration is many orders of magnitude higher than in water bodies, so as to increase production per volume. In these conditions, large-size light-harvesting antenna, an advantage in the wild, becomes detrimental for overall biomass productivity because high optical density of antenna pigments results in self-shading and light attenuation in deep layers of the culture. Incident light energy is mostly absorbed by cells at the surface, exceeding photosynthesis maximum rate and resulting in dissipation and loss of excess photons as

fluorescence or heat. Up to 80% of the absorbed photons could thus be wasted [33]. Moreover, these cells would be more subjected to photoinhibition, that is the light-induced inactivation of photosystem II due to photooxidative damage [42, 43], leading to losses in photosynthetic productivity [44]. Meanwhile, a suboptimal illumination would occur in the deepest layers, where energy consumption by respiration would reduce the overall yield [45, 46]. Only the intermediate layers are in conditions for optimal photosynthetic yield and thus determine the productivity of the mass culture. As a consequence of the light gradient formed across the photobioreactor diameter, the real solar radiation to biomass conversion efficiency decreases far below the calculated value of 8–10% [25], as also mentioned in the “[Present Algal Productivity in the Laboratory Versus Large Scale](#)” section. The unequal light distribution has another negative consequence: algal cells are stirred into the reactor where they can move from suboptimal illumination to strong light within few seconds, without having time to adjust the photosynthetic apparatus [47]. Intermittent light is highly stressful and contributes to the oxidative stress and photosystem II photoinhibition.

It was first postulated 60 years ago [48] that a truncated light-harvesting antenna size would optimize light utilization efficiency of microalgae in a mass culture. Such configuration would require a higher light intensity to reach photosynthesis saturation (refer to “[Present Algal Productivity in the Laboratory Versus Large Scale](#)” section for definition of light saturation constant) and would minimize wasteful dissipation of excess absorbed energy and photoinhibition. In particular, a better transmittance of light deeper into the culture would be allowed, enabling more cells to contribute to useful photosynthesis and culture productivity. It was estimated that a reduced optical density could improve solar radiation to biomass conversion efficiency up to three to four times [33]. [Figure 2](#) schematically compares the behavior of the wild type to a truncated antenna strain, with respect to light distribution and growth inside a hypothetical tubular photobioreactor.

While algae with a truncated light-harvesting chlorophyll antenna size would be useful in controlled mass culture conditions, they are not competitive and do not survive in the wild, so are not encountered in nature. Although the advantages of strains with altered optical

properties were recognized long ago, only recently, with molecular genetics coupled to biophysical phenotype screening methods, the problem of their construction can be addressed. Small antenna mutants can now be obtained through chemical/UV mutagenesis or insertion mutagenesis and tested for their light use efficiency. Small antenna strains described earlier carried mutations in pigment biosynthesis. Pigments are bound to photosystem protein subunits where they act cooperatively in light harvesting, meanwhile photosystem subunits require a specific set of bound pigments to be properly folded and assembled. Reduction of photosystem II antenna size was indeed observed as a consequence of mutations affecting the biosynthesis of chlorophyll *b* [49, 50] or the biosynthesis of xanthophylls [51], specifically coordinated by light-harvesting antenna subunits, in the model unicellular green alga *Chlamydomonas reinhardtii*. However, photosynthetic antenna components are devoted not only to light harvesting but also to photoprotection, meant as the ability to prevent reactive oxygen species in the light [52]. Moreover, carotenoids of the xanthophyll cycle have multiple roles in photoprotection [53–55]. Algal mechanisms of oxidative stress resistance are particularly important for growth in photobioreactors, especially the capacity to respond to fast alterations in light intensity that cells can face during stirring of the medium. Therefore, efforts at reducing the antenna must not compromise the photoprotection ability of the strain, otherwise the advantage obtained in terms of light distribution would be worthless. For example, pigments mutants of the diatom *Cyclotella*, without fucoxanthin, a xanthophyll synthesized in chlorophyll *a* and chlorophyll *c* containing algae, displayed a truncated antenna size and a higher light saturation, but no improvements in biomass productivities were observed in either laboratory cultures or outdoor ponds [29]. Every isolated antenna mutant must be tested for its growth performance and oxidative stress resistance before any application. Novel genes need to be found, whose deletion or mutation leads to phenotypes of selective downregulation of antenna components, retaining gene products with photoprotective function. Among available methods for the identification of these genes, random insertion mutagenesis is of choice (refer to “[Future Perspectives and Technological Developments](#)” section for further explanation).



**Algae, a New Biomass Resource. Figure 2**

Schematic representation of a tubular photobioreactor transversal section. In wild type (*left*), most of light is absorbed by cells in surface layers. Excess light is transformed into heat by physiologic dissipation mechanisms. An algal strain with truncated antenna (*right*) is less efficient in absorbing light, thus allowing penetration of the irradiance deeper into the culture. As a consequence, a higher fraction of cells is photosynthetically active and accumulates biomass and a smaller fraction of the incident sunlight energy is dissipated into heat due to the lower number of photons intercepted by each photosystem. However, since antenna protein subunits, beside light harvesting, have important role in photoprotection preventing the formation of reactive oxygen species or scavenging them, it is essential that the reduction in the antenna system does not compromise photoprotection capacity of the strain in use

Once a mutant library has been generated by transformation with a cassette carrying a selectable marker gene, strains with the desired characteristics must be identified from the pool by high throughput methods. Truncated antenna size strains can be selected for a lower chlorophyll fluorescence yield and/or pale green appearance and/or lower accessory pigments versus chlorophyll *a* content with respect to wild type. The fluorescence screening is performed upon application of a saturating light pulse to dark-acclimated cells. Light excites chlorophylls connected to photosystem II which undergo a transition from an “open” state with low fluorescence to a “closed” state when all photosystem II primary acceptors are reduced and fluorescence level is maximal. The maximum fluorescence parameter depends on the number of chlorophylls bound to photosystem II and its value is therefore roughly indicative of photosystem II antenna size. Thus, mutants with reduced chlorophyll antenna are expected to show a phenotype with a lower yield of chlorophyll fluorescence with respect to the wild type [56]. At room temperature, chlorophyll fluorescence derives from photosystem II, since photosystem I fluorescence yield is negligible in this condition.

Photosystem I is more efficient in quenching antenna excitation because of a faster turnover rate of the photochemistry reactions and its fluorescence is detectable only at low temperature (77 K). For this reason, room temperature fluorescence intensity is indicative of the sole photosystem II antenna. Meanwhile, since antenna subunits bind different pigments with respect to photosystem core complexes, namely chlorophyll *b* and xanthophylls, analysis of pigments by absorption spectra and/or HPLC can detect mutations in both photosystem I and photosystem II antenna systems. Since chlorophyll biosynthesis in the chloroplast is strictly coordinated with the expression of both nucleus and chloroplast-encoded chlorophyll-binding proteins, a “pale green” phenotype could arise from mutations affecting directly chlorophyll biosynthesis or from mutations impairing expression of chlorophyll-binding proteins, their import into the chloroplast or their assembly into photosystems. Mutations affecting these mechanisms are expected to yield strains with constitutively reduced antenna size, irrespective of the compensatory acclimative mechanisms which are involved in adapting the light-harvesting function to environmental conditions.

Factors controlling antenna proteins expression at the posttranscriptional level are attractive for genetic engineering manipulations [57] and a permanently active variant of NAB1, a novel cytosolic RNA binding protein, allowed to reduce photosystem II antenna size by 10–17% via translation repression, finally improving light-to-biomass conversion efficiency in high light. In particular, cell culture densities of the antenna mutant in the 200 mL bottles increased within 38 h from  $5.97 \times 10^5$  to  $1.2 \times 10^7$  cells mL<sup>-1</sup>, whereas the parental strain reached only  $5.72 \times 10^6$  cells mL<sup>-1</sup> after the same time period [58].

Coordinated downregulation of the entire light-harvesting chlorophyll-binding *lh*c gene family was achieved by RNA interference technology, at last resulting in increased efficiency of cell cultivation under elevated light conditions of 1,000  $\mu$ mol photons m<sup>-2</sup> s<sup>-1</sup>. Peak density of RNAi cultures was detected already after 26.5 h, at which parental strain cultures had only been able to grow to 54% of their maximal cell densities [59]. However, RNA interference-induced phenotypes can become unstable after some time. This technology would be useful to identify antenna components or modulators that can be depleted without compromising photoprotection ability. It should be noted, however, that strains for industrial use should be stable over time making deletion/insertional mutants preferable.

Through insertional mutagenesis of *C. reinhardtii*, an interesting mutant, called *tla1*, was isolated for a reduced yield of chlorophyll fluorescence. It displayed a residual chlorophyll content per cell of about 35–40% the wild-type level and a functional chlorophyll antenna size of both photosystem II and photosystem I reduced to 114 and 160 chlorophyll molecules, respectively [56]. This mutation allowed to identify a novel gene involved in the regulation of antenna proteins accumulation in algae. The *tla1* strain required a higher light intensity for the saturation of photosynthesis, about twofolds higher than wild type, and showed a higher photosynthetic productivity under mass culture conditions, reaching higher cell densities ( $10 \times 10^6$  cells mL<sup>-1</sup> vs.  $6.4 \times 10^6$  cells mL<sup>-1</sup>) [56, 60]. Simultaneous reduction of photosystem II and photosystem I antenna systems is a desired advantage in terms of global optical density of the cell. In contrast, the sole reduction in photosystem II antenna, if not

counterbalanced by an adjustment in photosystem II/photosystem I ratio, could decrease the photon use efficiency possibly because a portion of the light energy absorbed by photosystem I could not be efficiently utilized in the linear photosynthetic electron transport process [49]. The minimum number of chlorophyll molecules, needed for the assembly of the photosystem core complexes, has been estimated to be 37 for photosystem II and 95 for photosystem I [61]. This would be the minimum chlorophyll antenna size that would allow for assembly of the photosystems and that could be theoretically achieved through mutagenesis and genetic engineering [25]. Reduction in the antenna systems acquired so far is still above this estimated minimum chlorophyll antenna size and further improvements could be achieved in the future.

Although truncation of the antenna size of the photosystem unit can contribute to improve optical properties of the algal suspension, considering the important role of the antenna systems also in photoprotection, an alternative route to pursue could be to reduce the number of photosystems per cell and thus the pigment content per cell. In algae, acclimation to high light could rely more on the reduction of the photosystem density rather than on major adjustments in the antenna system to prevent overexcitation [62]. This natural light acclimation ability could be exploited in order to generate strains with a constitutive reduced optical density through the identification of involved regulatory factors and their genetic manipulation.

### Non-Photochemical Energy Quenching at the Molecular Level

Oxygenic photosynthesis involves highly reactive intermediates, such as singlet excited state of chlorophylls that can decay into triplet excited state, and may lead to the formation of reactive oxygen species (ROS). These by-products can damage the photosynthetic apparatus and other chloroplast constituents. The potential for damage is exacerbated when the amount of absorbed light exceeds the capacity for light energy utilization in photosynthesis, a condition that can lead to decreases in photosynthetic efficiency. Non-photochemical quenching (NPQ) is the process of heat dissipation of excess absorbed energy, which is aimed at regulating light harvesting and preventing



overexcitation of reaction centers. The major component of NPQ, called qE, is the fastest established and more rapidly reversible quenching deriving from de-excitation of the singlet excited state of chlorophyll in the light-harvesting antenna of photosystem II in response to a change in thylakoid lumen pH. Excess light with respect to the capacity of CO<sub>2</sub> fixation promotes lumen acidification since ATP synthesis is limited by the low ADP + Pi concentration available. In addition to qE, a more slowly relaxing component of the NPQ process is known as qI, from photoinhibition, that is the light-induced inactivation of photosystem II reaction center. More recently, qI was shown to depend on the synthesis of zeaxanthin, promoted in excess light. Zeaxanthin binds to light-harvesting antenna proteins where it becomes a quencher for chlorophyll excited states. Its action relaxes when zeaxanthin is released and back-transformed into violaxanthin [63]. A third quenching component, relaxing within minutes, has been reported and called qT and proposed to depend on the phosphorylation of the outer antenna proteins that controls their partition between photosystem II and photosystem I during the process of state 1–2 transition [64].

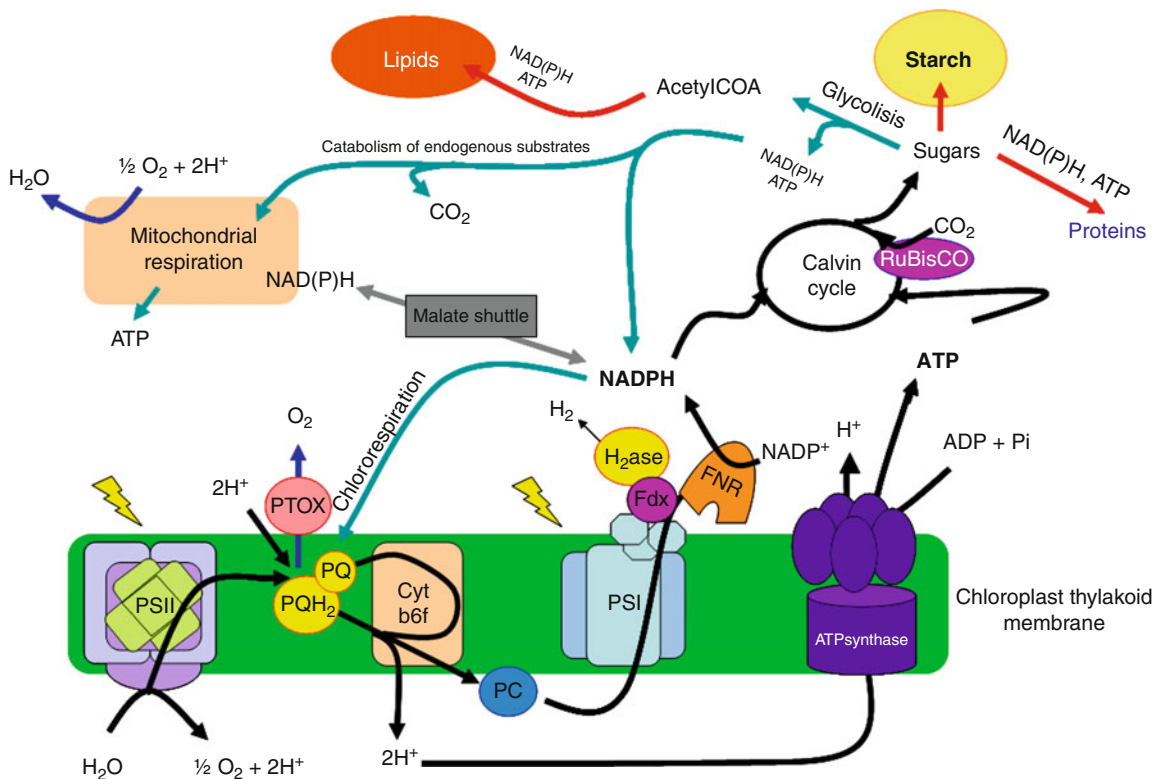
Photosynthetic light harvesting in higher plants is mainly regulated through qE and requires PsbS, a photosystem II subunit [65], which is sensitive to pH through two lumen-exposed glutamate residues whose protonation is required for qE activation [66]. Since PsbS does not bind pigments, it cannot be a quenching site but rather a pH sensitive trigger [67] transducing a conformational change to antenna proteins binding chlorophylls and xanthophylls. Changes in the mutual distance/orientation of these chromophores reversibly promote energy-dissipation processes such as transient formation of carotenoid radical cations, which decay to the ground state dissipating energy as heat [68, 69]. Although light-dependent energy quenching is a property of all photosynthetic eukaryotes, strong differences in the underlying mechanisms are apparent. Although genes encoding PsbS are found throughout the green lineage, indicating that the protein was present before the separation between unicellular green algae and multicellular organisms, a PsbS protein is only accumulated in green macroalgae and land plants. Nevertheless, several algal species, such as *Chlorella*

*zofingiensis* and *Scenedesmus communis*, exhibit strong light-dependent NPQ independent from light intensity during growth [70] or upon acclimation to high light as in *Chlamydomonas reinhardtii* [71], suggesting other gene products might be involved. The moss *Physcomitrella patens* and algal genomes, but not red algae and cyanobacteria, include LHCSR (stress-related members of the Lhc protein superfamily) genes that are essential for NPQ [72, 73]. Cyanobacteria and possibly red algae, which have phycobilisomes for light harvesting, perform excess energy dissipation by a mechanism distinct from qE, which depends on the Orange Carotenoid Protein (OCP) [74]. Understanding the mechanisms of NPQ and identification of the genes involved is valuable as a target for genetic improvement of algal productivity and domestication, since NPQ controls light stress resistance and energy losses in an algal mass culture.

### Interconversion and Storage of Photosynthetic Metabolic Products

Aside from light harvesting, solar-to-biomass conversion efficiency strongly depends on downhill biochemical reactions influencing photosynthate production and utilization. Respiration and other metabolic activities are accounted for about 30% energy losses with respect to the light energy absorbed leading to an estimated solar-to-biomass conversion efficiency of 8–10% in algae [25]. Biomass accumulation ultimately depends on algal photosynthetic architecture and the intricate relationships between oxygenic photosynthesis, mitochondrial respiration, and catabolism of endogenous substrates. In order to improve solar-to-biomass conversion efficiency, it is essential to recognize bottlenecks in the processes interconnecting photosynthesis to other metabolic pathways to make them targets for engineering. Figure 3 is a schematic representation of the main anabolic and catabolic pathways influencing photosynthate production and biomass accumulation.

The primary reactions of photosynthesis in algae and plants occur in the chloroplast thylakoid membrane and are catalyzed by the multiprotein-pigment complexes photosystem II (PSII), cytochrome *b6f* (Cyt *b6f*), and photosystem I (PSI). The photosynthetic apparatus itself is a very flexible system, able to modify



**Algae, a New Biomass Resource. Figure 3**

Metabolic overview. Components of the photosynthetic electron transport chain embedded in the chloroplast thylakoid membrane are depicted. *Black lines mark biosynthetic reactions building up the main constituents of the cell, proteins, starch, and lipids. Dark gray lines mark catabolic reactions that consume endogenous substrates.* PSII photosystem II, PSI photosystem I, Cyt b6f cytochrome b6f complex, PQ plastoquinone, PQH<sub>2</sub> plastoquinol, PTOX plastid terminal oxidase, PC plastocyanin, Fdx ferredoxin, FNR ferredoxin-NADP<sup>+</sup> oxidoreductase, H<sub>2</sub>ase hydrogenase, RuBisCO ribulose-1,5-bisphosphate carboxylase/oxygenase

exciton fluxes within its antenna complexes as well as electron fluxes among its electron transfer components [75]. Photosystem II and photosystem I operate in series within the photosynthetic electron transport chain. However, alternative electron transport pathways could be engaged, as well as alternative electron sources or sinks that ultimately would affect availability of reducing power (NADPH) and ATP for carbon dioxide fixation. As a matter of fact, the net products of photochemistry, NADPH and ATP, that supply the Calvin–Benson cycle, are also used by other metabolic pathways such as nitrate assimilation, lipid, aminoacids, and pigments synthesis. These different sinks may significantly contribute to modify the ATP/NADPH ratio and, as a consequence, in vivo carbon

fixation [76]. The photosynthetic process is divided into reactions that provide ATP and NADPH (photochemical light reactions) and reactions that consume both compounds (carbon assimilation through the Calvin–Benson cycle). The rate of these two phases may differ in several orders of magnitude, especially in high light when carbon assimilation is limiting the overall process, and alternative electron sinks may have an important photoprotective role to consume excess generated reductants. Respiration in mitochondria can serve as well as a sink for excess photosynthesis generated reducing power. Complex interactions between photosynthesis in the chloroplast and respiration in the mitochondria occur because both processes are linked by common key metabolites such as

ADP/ATP, NAD(P)H, triose-P, and hexose-P [75, 77, 78]. The chloroplast continuously communicates its metabolic state to the cell, for instance via the export of carbohydrates, which are synthesized during photosynthesis. The transport of carbohydrates across the chloroplast membrane directly allows to exchange reducing equivalents through “redox valves,” such as the malate-oxaloacetate shuttle [79]. Different cellular compartments are thus strictly interconnected and photosynthesis is influenced by either the requirements of the chloroplast or the metabolic status of the whole cell. In particular, it was observed that efficient photosynthesis depends also on mitochondrial respiration in the light [16, 78, 80, 81].

Carbon fixation is influenced by alternative electron sinks as well as alternative electron sources. For example, reducing equivalents derived from catabolism of endogenous substrates can fuel the respiration chain in mitochondria but also the photosynthetic electron transport chain, thanks to a respiration pathway in the chloroplast. Respiration of the chloroplast, also called chlororespiration, was defined as a respiratory electron transport chain in interaction with the photosynthetic chain [82] and could have its origin in the cyanobacterial endosymbiotic ancestor of chloroplasts [82, 83]. This pathway was discovered about 30 years ago in algae and originally proposed to account for the changes in the redox state of plastoquinone, a photosynthetic electron carrier, in darkness [84]. Chlororespiration includes a NAD(P)H-dehydrogenase [85, 86] and a plastid terminal oxidase PTOX [87–91]. The NAD(P)H-dehydrogenase enzyme of chlororespiration is possibly involved also in recycling photosynthesis-generated NAD(P)H, establishing a cyclic electron transport around photosystem I [92–94]. This alternative electron transport could be engaged to balance the ATP/NADPH stoichiometry, since the net product is only ATP, in response to the requirements of the cell and is controlled by the redox state of the plastoquinone pool. The strong interplay between mitochondria and chloroplast metabolism is also evident from the characterization of *Chlamydomonas reinhardtii* mutants with defects in the mitochondrial electron transport chain: an enhanced glycolysis, to compensate for the absence of mitochondrial ATP input, is thought to increase plastoquinone reduction through chlororespiration

and favor cyclic over linear electron transport in the chloroplast [80].

Cell fitness depends on the delicate equilibrium between biosynthetic and catabolic reactions, both essential. In order to use a photosynthetic organism as a biofuel or biomass producer, it is essential to control these intricate metabolic dynamics. Carbon assimilation rate may be a bottleneck, particularly in high light. *Dunaliella salina* was shown to upregulate key enzymes in the Calvin cycle at high salinity, possibly in order to enhance synthesis of glycerol as osmotic protector [95]. It is conceivable that improvements in carbon assimilation could be achieved in the future with the help of genetic engineering. Strategies to achieve this aim could be altering expression and/or activity of key enzymes or acting on ATP/NADPH availability.

### Screening for State Transition as an Indirect Mean to Select Strains with Altered Redox Metabolism

The redox state of the plastoquinone pool is known to control a process called “state1–2 transition.” Since the light-harvesting antennae of photosystem II and photosystem I have distinct absorption spectra, changes in the spectral composition of the incident light can lead to unequal excitation of the two photosystems and thus to a decreased photosynthetic yield. Plants and algae are able to balance the relative excitation of the photosystems through state transition, that is the reversible migration of a fraction of the light-harvesting antenna of photosystem II from photosystem II in state 1 to photosystem I in state 2, upon phosphorylation [96–98]. This process was discovered about 40 years ago in unicellular microalgae [99,100]. The state transition process has been widely studied in the unicellular green alga *C. reinhardtii*, where its amplitude is larger than in plants and appears to have a major role in balancing the ATP/NADPH stoichiometry by regulating the switch between linear and cyclic electron flow, in addition to the redistribution of the excitation energy between photosystems following changes in light conditions. Indeed, recycling of electrons around photosystem I was observed in state 2 [92–94]. The interplay between chloroplast and mitochondria metabolism is strong in algae. The mere inhibition of

state transition was shown to be insufficient to modify photosynthesis in the presence of active mitochondrial respiration, in contrast it is essential when respiration is compromised, revealing the physiological significance of state transition in the energetic contribution [81]. Up to 80% of the excitation energy absorbed by the photosystem II antenna can be redistributed from photosystem II to photosystem I in *C. reinhardtii* that is therefore well suitable for screening state transition mutants [96, 97, 101]. Large chlorophyll fluorescence changes occur during a transition from state 1 to state 2, due to the fact that most of the photosystem II antenna is displaced from photosystem II to photosystem I during this process and chlorophyll fluorescence at room temperature arises mostly from photosystem II. The difference in fluorescence between states 1 and 2 can be measured with a fluorescence video imaging system and used for screening mutants deficient in state transition. In the model plant *A. thaliana*, states 1 and 2 are induced by illuminating cells with light preferentially absorbed by photosystem I and photosystem II, respectively [102]. In *C. reinhardtii* cells, however, this procedure is not effective because in this alga the absorption spectra of photosystem II and photosystem I antennae overlap. Thus, states 1 and 2 are achieved in the dark by taking advantage of the ability of the chlororespiratory chain to change the redox state of the plastoquinone pool [96, 97]. State transition and cyclic electron transport around photosystem I are controlled by the redox state of the plastoquinone pool. This diffusible electron transport element is at the crossroad between linear electron transport from water to  $\text{NADP}^+$  and the chlororespiration pathway dissipating excess redox power by reducing  $\text{O}_2$  to  $\text{H}_2\text{O}$ . Thus, plastoquinone redox state is altered either by an imbalanced excitation of the photosystems or by changes in the rate of catabolic degradation of storage molecules and could therefore serve as a key sensor of both the incident photon flux and the cellular energetic status. Mutations in state transition can affect the non-photochemical reduction and oxidation of the plastoquinone, catabolism of endogenous substrates, and upstream reactions involved in feeding reducing equivalents into the stroma compartment of the chloroplast. Screening for state transition mutants is a useful strategy to isolate strains with an altered redox metabolism that could finally affect photosynthate

generation/storage, growth rate, and biofuel production. It represents a mean of domestication alone or in combination with other characteristics.

### Accumulation of Biomass as Neutral Lipids

Assimilated carbons can be stored in high energy value metabolites, such as starch, proteins, or lipids, as shown in Fig. 2. Some species of algae are particularly attractive because of their ability to produce high amounts (20–50% of dry cell weight) of neutral lipids that can serve as a source of biodiesel. Cellular lipid metabolism is altered toward the accumulation of neutral lipids, mainly triacylglycerols, under stressful conditions, such as nutrient starvation, salinity, nonoptimal growth medium pH, low temperature, high light, but also during aging of the culture [2]. These lipids do not have a structural role in membranes but serve as a storage form of carbon and energy, confined in lipid bodies in the cytoplasm. Lipid bodies also occur in the inter-thylakoid space of the chloroplast in certain green algae such as *Dunaliella bardawil* [103]. An attractive oleaginous green alga is *Botryococcus braunii* that produces up to 80% of dry weight of very-long-chain (C23–C40) hydrocarbons similar to those found in petroleum, under adverse environmental conditions [8, 104]. Neutral lipids can serve additional physiological roles. Fatty acids synthesis consumes twice the NADPH required for carbohydrate or protein synthesis and may thus provide an electron sink under photooxidative stress [2]. Moreover, coordination with carotenoid synthesis has been observed, in particular carotenoids are sequestered into cytosolic lipid bodies where they function as a sunscreen to reduce light striking the chloroplast [2, 105, 106].

Algae triacylglycerols can be exploited to produce biodiesel, via esterification of fatty acids. Biodiesel has a three to four times higher energy yield than ethanol; however, biodiesel still represents a small percentage of total diesel fuel consumption (1.6% in Europe and 0.21% in the USA, 2005–2007), while ethanol represents 5% of US gasoline consumption [23]. Currently, biodiesel is mainly produced from higher plants, such as palm and soybean [23]. But while extensive cultivation of crops to supply energetic demands is unsustainable, competing with food industry for arable

lands, oil production per hectare from algae, based on theoretical calculations, would be 100-fold higher than that of soybean and could meet 50% of present US transportation demand using less than 3% of available cropland [9]. Based on experimentally demonstrated biomass productivity, oil yield of microalgae could be 136 or 58 t ha<sup>-1</sup> year<sup>-1</sup> considering 70% or 30% oil in biomass, respectively [9]. However, insufficient efforts for the establishment of algae-based biodiesel production plants have been made until now. The ability of algae to produce neutral lipids is species-specific rather than genus-specific [2]. Interesting species for biodiesel production include green algae, such as *Botryococcus braunii*, *Neochloris oleoabundans*, *Nannochloris* sp., *Chlorella* sp., and *Dunaliella primolecta*; diatoms, such as *Nitzschia*, *Phaeodactylum tricornutum*, and *Cylindrotheca* sp.; and members of other algal taxa, such as *Nannochloropsis* sp. and *Schizochytrium* sp. [9]. Relatively few algal strains have been examined to date, among the species available in nature. The possibility to isolate new oleaginous strains with the ability to accumulate high levels of oils with the best properties as fuel is likely to offer substantial improvements in lipid yield of industrial cultures while mutagenesis will further enhance the productivity of natural strains. Fatty acid composition can vary both quantitatively and qualitatively with the physiological status and culture conditions. Since oil accumulation is enhanced under stress, altering growth medium composition appears as a strategy to improve oil productivity. For example, high carbon dioxide concentrations (>5% v/v) in *Dunaliella salina* [107] or nutrient deficiency such as nitrogen deprivation in *Chlorella vulgaris* [108] can lead to a threefold increase in intracellular lipids.

Lipid metabolism has been poorly studied in algae. Nevertheless, available data suggest that the basic pathways are analogous to those experimentally detailed in higher plants. However, differences are distinguishable:

1. While in higher plants organic carbon is translocated from photosynthetically active tissues to sinks for lipid synthesis and storage, in microalgae triacylglycerols accumulation takes place within a single cell together with photosynthesis.
2. While neutral lipids synthesis is mainly associated to seed development in higher plants, it is triggered under stress conditions in algae [2].

A better knowledge of lipids synthesis pathways and regulation mechanisms is needed in order to implement genetic engineering strategies for oil production. Identification of metabolic differences between oleaginous strains and the species that do not accumulate substantial amounts of lipids is a possible research strategy to be realized using comparative transcriptomic, proteomic, and metabolomic profiles of different strains or the same strain under control versus stressful conditions. Microarray analysis of *Chlamydomonas reinhardtii* transcripts under anaerobic incubation has been performed [109], revealing fermentative pathways that produce acetyl CoA, the substrate for fatty acids synthesis. Genes encoding enzymes involved in these fermentative reactions represent putative candidates for increasing triacylglycerols accumulation [2].

The limiting step in fatty acids biosynthesis is the reaction catalyzed by the first enzyme of the pathway: acetyl CoA carboxylase (ACCase). The properties of this enzyme have been characterized in *Cyclotella cryptica* [110–113], but attempts to alter its expression level did not have effects on lipid production [113]. Some transformants showed two- to threefold higher ACCase expression and activity than wild-type cells; however, no detectable increase in lipid levels was observed. Overexpression of the endogenous enzyme may induce negative feedback so that increased activity of the ACCase enzyme could be compensated for by other pathways within the cell [113]. Indeed, feedback inhibition was reported in higher plants [114], where only a heterologous enzyme was successfully overexpressed really improving oil content [115, 116]. This approach could be proposed again and plants genes could be used for overexpression in algae.

Although activity and expression of single enzymes could be altered, a more powerful strategy might consist in the identification of transcription factors co-regulating all the genes of the pathway.

Lipids biosynthesis is strictly interconnected to other biosynthetic pathways in a unique global network. A *Chlamydomonas reinhardtii* starch-less mutant, defective in the ADP-glucose pyrophosphorylase, was shown to hyper-accumulate triacylglycerol by a factor of 10 [117]. A strategy to improve oil synthesis in order to produce biodiesel could be as well to act on the partitioning of carbon and energy between different pathways.



Finally, fatty acids composition of triacylglycerols influences diesel quality and must be considered for commercial application. Genetic engineering can additionally serve to alter the fatty acids profile thus improving fuel properties in engines [23]. Many algae can synthesize very long polyunsaturated fatty acids in large amounts (arachidonic C20:4, eicosapentaenoic C20:5, docosahexaenoic C22:6 acids), that are extracted as high added value products. Polyunsaturated fatty acids decrease in nutrient-limited medium and during heterotrophic growth, concomitantly with the increase of C16:0 and C18:1 fatty acids in triacylglycerols [118–120]. Monounsaturated fatty acids are indeed suitable for good diesel properties [23].

Research of new genes involved in lipid metabolism could be pursued by random mutagenesis. A mutant library represents a biological resource of novel strains that could include improved biodiesel producers. An easy and rapid strategy to screen for mutants with altered lipid content is by the use of Nile red [121]. Nile red is a hydrophobic molecule that emits a significant fluorescence signal when dissolved in lipids at 565–585 nm, in a spectral region where photosynthetic pigments fluorescence is negligible. Nile red fluorescence detection could be applied to screen and compare mutant colonies with respect to

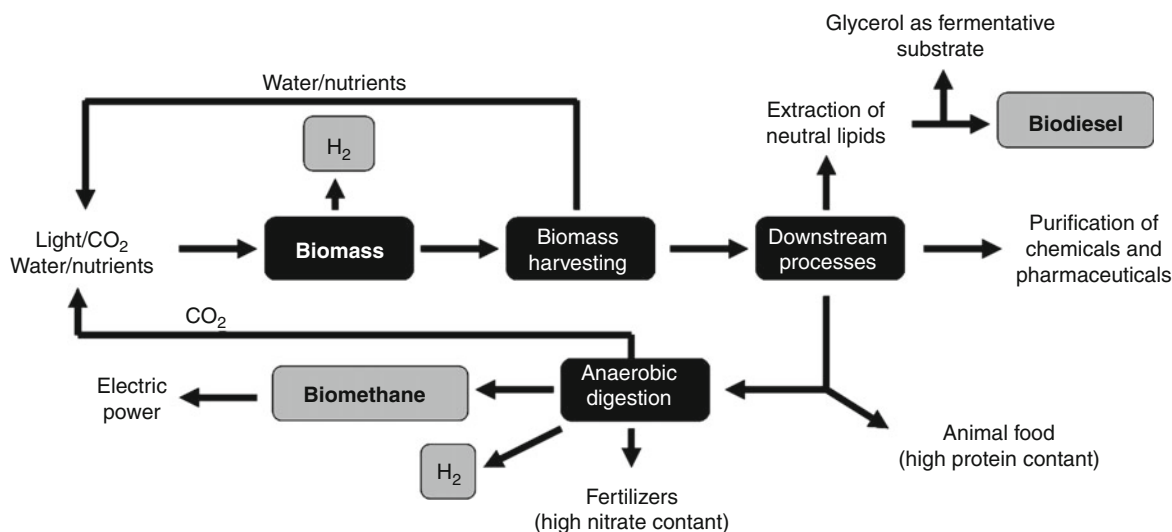
the wild type in a microtiter plate. Mutants displaying an altered Nile red fluorescence yield and thus altered lipid content with respect to the wild type can then be studied by molecular biology techniques to identify the mutated gene and tested for oil productivity.

### Planning an Algal Refinery

Algae convert solar energy into chemical energy and produce biomass through photosynthesis. The process requires light, carbon dioxide, water, and other essential nutrients and can be considered renewable. Minimal nutritional requirements can be estimated using the approximate molecular formula of the microalgal biomass, that is  $\text{CO}_{0.48} \text{H}_{1.83} \text{N}_{0.11} \text{P}_{0.01}$  [9, 122]. However, inorganic elements have to be added in excess because they form complexes in solution that are not bioavailable.

Algal biomass can be exploited in different ways, providing several products and biofuels. A scheme of the possible organization of an algal refinery is reported in Fig. 4.

Upon harvesting of the biomass, water and nutrients can be recycled, especially using closed photobioreactors. Then, biodiesel is obtained from



**Algae, a New Biomass Resource. Figure 4**

Scheme of the possible utilizations and products of an algal refinery

esterification of fatty acids contained in triacylglycerols or from other neutral hydrocarbons accumulated by oleaginous algae. A by-product from esterification of fatty acids to produce biodiesel is glycerol that can serve as fermentative substrate. For example, alga *Schizochytrium limacinum* produces docosahexaenoic acid, a polyunsaturated  $\omega 3$  fatty acid with beneficial effects for human health, through fermentation on crude glycerol [123]. Currently, algae are already commercially used to extract chemicals and pharmaceuticals and this utilization can be combined with the need for biofuels. High added value products derived from algae include polyunsaturated fatty acids, proteins, biopolymers, polysaccharides, pigments, vitamins, and antioxidants [1, 4]. To date, commercialized algae are not transgenic, but this scenario could change in the near future. For instance, algae can be transformed in “cell factories” with the help of molecular biology techniques for manufacturing recombinant proteins with pharmaceutical or other applications, an approach also called “molecular farming.” For example, expression of human antibodies and vaccines has already proven to be successful [124–128]. In general, algae are attractive as expression systems for the short life cycle, the cheap and easy large-scale cultivation, and the ability to grow up to high cell densities [129]. Algae have been shown free of human pathogens and toxins and thus considered safe.

After extraction of oils and other products, the spent biomass can serve as animal food or used as fermentative substrate for other microorganisms whose metabolic by-products include biomethane and biohydrogen. Effluents from the digestion can be used as fertilizers because of their high content in nitrate. Biomethane (or biogas) has a high energetic yield and is produced by anaerobic bacteria during fermentation of the biomass. The obtained gas contains methane (50–75%) but also carbon dioxide (25–50%). However, released carbon dioxide is equivalent to the carbon that algae have assimilated during growth, leading to zero net carbon dioxide emissions. Carbon dioxide released during the process can be immediately recycled for alga growth. Electric power produced from the biogas can sustain the energy demands of the photobioreactor plant itself [10]. Some algae have been evaluated for biomass conversion to methane, in particular the giant kelp *Macrocystis pyrifera* because of

its high growth rate and ease of harvesting [5]. Alternatively, algal biomass can serve as fermentative substrate for heterotrophic or photosynthetic anoxygenic sulfur bacteria that use electrons coming from organic carbons to produce molecular hydrogen. In particular, photosynthetic anoxygenic sulfur bacteria use light energy to drive a fermentative reaction and organic acids are electron donors for molecular hydrogen evolution. Generated hydrogen is a high energetic fuel and clean since its combustion generates water [130].

In addition, algae have directly been evaluated for the production of biohydrogen. As a matter of fact, some algae are able to evolve hydrogen by a hydrogenase enzyme that catalyzes reduction of protons to molecular hydrogen. Hydrogen metabolism is primarily the domain of bacteria and microalgae. Microbial hydrogen formation is catalyzed by either nitrogenases or hydrogenases, enzymes that can only function under anaerobic conditions. Nitrogenases are used by certain cyanobacteria and photosynthetic bacteria, whereas green algae use hydrogenases [12]. In *Chlamydomonas reinhardtii*, the hydrogenase is expressed in anoxia together with the enzymes of the fermentative metabolism that is active in the dark [109]. However, algae are attractive for hydrogen production if photosynthetic electron transport can be directly exploited, since hydrogen would be generated from the most abundant of the natural resources, sunlight and water. Interestingly, after a period of anaerobic incubation in the dark, photosynthetic hydrogen evolution is detected transiently upon illumination [131]. The hydrogenase is transiently active during the dark-to-light transition, to consume reducing power in the time needed to fully activate the Calvin–Benson cycle, and receives electrons from ferredoxin reduced by photosystem I. This activity lasts until oxygen evolution by photosystem II restores aerobiosis and inhibits the hydrogenase enzyme. In order to photo-produce hydrogen on large scale, the major challenge to overcome is the inhibition by oxygen evolved by photosystem II during oxygenic photosynthesis. Up to now, encouraging hydrogen production rates have been obtained by a two-stage approach: the severe oxygen sensitivity of the hydrogenase is circumvented by temporally separating photosynthetic oxygen evolution and growth (stage 1) from hydrogen evolution (stage 2). A transition from stage 1 to stage 2 is performed upon sulfur

deprivation, which reversibly inactivates photosystem II and oxygen evolution. Under this condition, oxidative respiration by the cell in the light depletes oxygen and causes anaerobiosis in the culture, which is necessary and sufficient for the induction of the hydrogenase [13, 15]. Electrons for hydrogen production originate from the residual activity of photosystem II as well as from the consumption of endogenous substrates that generates extra electrons supplied to the photosynthetic chain through the chlororespiration pathway [16, 86]. Hydrogen production depends from available electrons sources and competitive sinks of reducing power and a strategy to improve its evolution would also be to increase electrons channeling toward the hydrogenase and to deplete competitive electron sinks. An improved hydrogen evolution under sulfur starvation was observed in a mutant affected in the mitochondrial respiratory chain and simultaneously impaired in the ability to activate state transition [16]. In this mutant, accumulation of starch in the chloroplast supplied the photosynthetic electron transport chain through chlororespiration, while downregulation of cyclic electron transport around photosystem I enabled a greater fraction of electrons to be used by the hydrogenase (see “[Interconversion and Storage of Photosynthetic Metabolic Products](#)” and “[Screening for State Transition as an Indirect Mean to Select Strains with Altered Metabolism](#)” sections for further explanations about state transition and cyclic electron transport). The hydrogen production rates of the mutant were 5–13 times higher than in wild-type strain, yielding about 540 mL of hydrogen per 1 L of culture over a 10–14-day period (up to 98% pure) [16]. However, sulfur deprivation is deleterious for the cell and this system is sustainable only for few days. Upon a recovery phase of sulfur repletion, and reconstitution of reserves as endogenous substrates, hydrogen production can be resumed. Nevertheless, continuous hydrogen production is not possible using this technology. An alternative two-stage approach to overcome the oxygen sensitivity of the hydrogenase, with respect to sulfur deprivation, involves the control of photosystem II expression [132]. The initial requirement for this application is a photosystem II-less mutant which would be complemented by an inducible cassette to switch on/off photosystem II activity and oxygen evolution. However, the most efficient method

to obtain large-scale hydrogen production is believed to be the continuous production of hydrogen concomitantly to photosynthesis and growth. The limiting factor for such a technology is the finding of a hydrogenase enzyme with sufficient resistance to oxygen. Two different approaches are possible: to engineer existing hydrogenases by introducing mutations conferring resistance to oxygen or to find organisms that can synthesize hydrogen even in the presence of some oxygen, thus having hydrogenases less oxygen sensitive. Such hydrogenase could then be heterologously expressed in algae [130] to generate a hydrogen producing strain with increased yield of this biofuel.

### Large-Scale Systems

Open ponds and closed photobioreactors have been evaluated for algae mass culture [9, 26, 30]. Currently, commercial exploitation of algae mainly utilizes open ponds, easier to construct and scale up, and less expensive. Open ponds can be categorized into natural waters (lakes, lagoons, ponds) and artificial ponds or containers. They are generally raceway cultivators driven by paddle wheels or unstirred. In contrast, closed systems like photobioreactors are more costly. Moreover, while many photobioreactors are operative at laboratory scale, their scaling up is more difficult and could be a limiting step. This is counterbalanced by several advantages over open ponds and some photobioreactors for large scale are in the process of being constructed. Closed systems are in the early stages of development but they are likely to become more attractive in the future.

1. Although photobioreactors could be artificially illuminated, to sustain growth during night, the use of natural sunlight is the major option for cost-effectiveness. Photobioreactors can be designed to improve the ratio between the illuminated surface and the total volume in order to minimize dark zones. From this point of view, flat-plate, horizontal and inclined tubular photobioreactors are promising [30]. A general tendency is to reduce the thickness, thus laminar layer plate or thin diameter tube configurations are becoming strongly favored, mimicking the leaves of land plants. For instance a 100-year-old, 10-m high lime tree shading an area of 100 m<sup>2</sup> has a leaf surface area of more than 2,500 m<sup>2</sup>. Expressed as

a surface/volume ratio, this amounts to a value of  $2.5 \text{ m}^2 \text{ m}^{-3}$  [26] and the design of large-scale photobioreactors should tend to this value.

2. Large-scale cultivation of algae requires first of all a good understanding of the physiology and ecology of the species, for example, with respect to nutrients requirements, optimum pH, and temperature. Close systems allow for a better control of growth conditions, with respect to pH, salts,  $\text{CO}_2$  partial pressure, temperature, mixing, and light supply.
3. Photobioreactors allow for a shorter period until net productivity is reached, approximately 2–4 weeks over 6–8 weeks in open ponds [26].
4. The stability of the dominating algal species is easier to be maintained in photobioreactors because contamination by other organisms is lower. The growth process, thus, is more reproducible and standardized. Strict sterility and quality control allowed by closed systems are particularly required for the production of pharmaceuticals and food additives [133].
5. In closed systems, water and nutrients, such as salts, can be recycled for algal growth and loss of water due to evaporation and diffusion of  $\text{CO}_2$  to the atmosphere would be minimized.
6. Efficient stirring mechanisms can be developed in photobioreactors ensuring homogeneity of the culture and finally influencing productivity in terms of growth. Flow is produced using either a mechanical pump or a gentler airlift pump. Mechanical pump could be more stressful and a strain with cell wall would be more suitable, because less sensitive to shear damage. Agitation is required to prevent cells sedimentation and to allow proper gas exchanges. As a consequence of the stirring and depending from the light gradient inside the culture due to the optical property of the strain, cells can suddenly be moved from suboptimal illumination to strong light at the surface without time to adjust the photosynthetic apparatus and finally leading to light stress. However, it was observed that light–dark cycling of above a certain frequency, such as 10 ms, can increase growth as compared with continuous illumination of the same intensity. It was proposed that a short dark period could allow the photosynthetic apparatus to recover from the excited state of the previous illumination pulse under light saturation conditions [9].
7. Harvesting of biomass is expensive since it requires energy. Biomass can be recovered from the broth by filtration or centrifugation [134]. Since concentration of biomass in photobioreactors is nearly 30 times higher than in open ponds ( $2\text{--}8 \text{ g L}^{-1}$  vs.  $0.1\text{--}0.2 \text{ g L}^{-1}$ ) [26], a much smaller volume of the photobioreactor broth needs to be processed to obtain a given quantity of biomass [9]. Once the biomass is harvested, processes to disrupt the cells and to extract desired products are being developed, for example, the high yield and scalable recovery of high purity fatty acids [135].

As a consequence of all aforementioned advantages of photobioreactors, productivity could be much higher than that of open systems and could balance the higher investment costs [9] (refer to “[Present Algal Productivity in the Laboratory Versus Large Scale](#)” section for biomass productivity data).

### Future Perspectives and Technological Developments

Mutagenesis and genetic modification of algae for bio-fuel production has received so far little attention, despite molecular biology procedures appearing essential in order to remove limitations in the economic sustainability of the biofuel production process. In particular, a biotechnological approach is required to obtain domestication of algal strains for cultivation in mass culture in order to:

1. Improve light distribution into a photobioreactor by decreasing the optical density of the single cell
2. Increase the light intensity at which photosynthesis saturates, thus reducing light energy dissipation and photoinhibition
3. Increase resistance to photooxidative stress
4. Optimize light utilization and carbon assimilation, finally improving solar-to-biomass conversion efficiency
5. Channel the metabolism toward oil accumulation for biodiesel production and to adjust the fatty acid composition to fit required fuel quality
6. Produce high value products and recombinant proteins of interest for pharmaceutical and chemical companies

When considering the advancement that biotechnology can bring to algae cultivation, the algal system is extremely attractive for commercial utilization if transformation and thus genetic improvement are possible. Nevertheless, the basic requirements for genetic improvement are so far available for a low number of species which are not necessarily suitable for the many possible applications. The basis for domestication of algal species is a good general knowledge of the biology for the target alga, including ecological, physiological, biochemical, and genetic aspects. In addition, easiness of cultivation in the laboratory and availability of a protocol for breeding are of paramount importance. On this basis, further improvement requires development of transformation protocols to introduce a foreign DNA into the cell and identification of markers for selection of transformants. Currently, about 25 species are accessible to genetic transformation, mainly green and red algae [1]. However, many algal species, although of interest for commercial use, cannot be propagated in the laboratory and cannot be transformed. Overcoming this limitation will be a major target of research in the future.

### Nuclear Genetic Transformation

Methods to introduce DNA into an algal cell exploited so far include vortexing in the presence of micro- or macroparticles such as glass beads, routinely used to transform *Chlamydomonas reinhardtii* [136] and successful for other green algae such as *Chlorella ellipsoidea* [137]; microparticle bombardment, used to transform species with a recalcitrant cell wall such as the diatoms *Phaeodactylum tricornutum* and *Cyclotella triptica* [138]; electroporation, effective in *Chlamydomonas reinhardtii* [139] and *Dunaliella salina* [124]. Cell wall-less mutants or protoplasts after cell wall digestion are more easily amenable to transformation. The limiting step is to allow survival of cells. Markers to select the transformants are classified as dominant, for example, conferring resistance to paromomycin [140] or zeomycin [136], or recessive. The latter require auxotrophic mutants and the corresponding endogenous intact gene for complementation. This strategy requires no exogenous gene expression [1]. As for genome sequence information, many genome sequencing projects, mainly for microalgae, are completed or in

progress, including green algae *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Volvox carteri*, *Dunaliella salina*, *Ostreococcus tauri*, and *Ostreococcus lucimarinus*; red algae *Cyanidioschyzon merolae*, *Porphyra yezoensis*, and *Porphyra purpurea*; and the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. Additionally, there has been an increase in expressed sequence tags (ESTs) databases.

Besides the methods for introduction of DNA into the cell, genetic transformation with a recombinant cassette requires regulatory elements active in the target species to drive transgene expression and availability of such regulatory elements is often limiting. Promoters can be heterologous, such as the cauliflower mosaic virus promoter CaMV35s that proved functional in several algae [1]; however, heterologous sequences are often not recognized by the host cell. Alternatively, endogenous promoters can be exploited. Endogenous promoters can be either constitutive, such as the promoter of the small chain of the ribulose biphosphate carboxylase, RBCS2 [136], the promoter of the Psad subunit of the photosystem I core complex [141], or the promoter of the heat shock protein 70A [142] in *Chlamydomonas reinhardtii*, or inducible, such as the promoter of the cytochrome C6 by copper deprivation in *Chlamydomonas reinhardtii* [132] or the arylsulfatase promoter inducible by sulfur deprivation in *Volvox carteri* [143]. Important regulatory elements can be included within introns and their addition may be useful for improving expression levels [144]. Once the transgene is introduced in the host genome, transgene silencing often occurs due to positional effects and action of epigenetic mechanisms [145], particularly active in the case of sequences located in the nucleus. Improvement of transgene expression is required for genetic engineering and for exploiting algae as “cell factory.” Inducible expression systems can be useful to partially overcome silencing of the recombinant gene.

### Chloroplast Transformation

An alternative to nuclear genome transformation is chloroplast transformation that has potentially many advantages:

1. Transgene expression is more stable and uniform between different lines, because the mechanism of homologous recombination active in the



chloroplast generates uniform clones, while epigenetic control, positional effects, and gene silencing are not present. Thus, known variability of transgene expression only derives from plastid differentiation stages.

2. High level of expression is favored by high polyploidy.
3. Multiple transgenes expression upon a single transformation event is possible due to the organization of genes into operons.
4. The plastid genome shows uniparental inheritance in many species, offering a degree of transgene containment.
5. *Chlamydomonas reinhardtii* has a single chloroplast, which contains approximately 80 copies of its own genome. Obtaining homoplasmy, where all copies of the plastid genome are transformed, is the main obstacle to the application of this technology to many plant species. In contrast, this process is easier in *Chlamydomonas reinhardtii* because of its single chloroplast and is achieved by a sufficient number of cells division under high selective pressure, making *Chlamydomonas reinhardtii* chloroplast a suitable factory for recombinant proteins.

Nevertheless, the influence of nuclear factors controlling plastid gene expression at multiple levels may limit transgene expression level in the plastid [133, 146]. A better comprehension of these mechanisms and of nucleus-plastid signaling is currently a major subject of research.

### Mutagenesis of the Nuclear Genome

Mutagenesis, besides its importance to investigate the physiological role of genes by analyzing phenotypic effects of their inactivation, is as well a powerful tool to generate strains with improved characteristics with respect to the wild type. Mutagenesis of the nuclear genome is the only way to obtain mutants for the majority of genes. Early work was performed using chemical or UV mutagenesis, a method that is of importance for species still not amenable to transformation. Chemical mutagenesis has evolved into a novel strategy of Targeted Induced Local Lesions in Genomes (TILLING) [147] where the chemical treatment is used to generate allelic series of point mutations. The target gene is then amplified by PCR and analyzed to verify

if mutations have occurred. TILLING is therefore a technique to isolate point mutations within a target gene that overcomes the “null phenotype” limitations of insertional mutagenesis and generates stable mutants. This technique has been so far applied to crop plants but not yet to algae and could be used as an approach of reverse genetics, aimed to investigate the physiological role of a gene with a known sequence. In applied research, TILLING could be pursued to obtain a stable mutant for a target gene whose implications for applicative purposes have been recognized. This approach is of choice if transgenesis should be avoided. However, limitations to this strategy are due to the high number of induced mutations per cell in order to increase the probability of obtaining mutations in the target gene. Several backcross cycles are then needed to isolate the desired mutation from all the others that decrease cell fitness. This lengthy procedure has so far discouraged application of the technique to algae and in fact makes it impossible in species where sexual reproduction is not available as a breeding procedure.

When dealing with species in which transformation procedures are available, insertional mutagenesis is generally preferred. Transforming the cell with an exogenous DNA vector helps to select the transformants, thanks to a selectable marker and to localize the mutation by identifying the genomic regions flanking the known insertion cassette. Integration of the exogenous DNA in the nuclear genome of eukaryotic algae occurs randomly, by non homologous recombination. Insertion of the cassette could take place within an encoding sequence, interrupting the gene and leading to a knockout, or it could as well perturb the promoter or other regulative sequences. This strategy can thus be used to generate random insertional libraries. The number of insertions per cell can be controlled by adjusting the DNA/cell ratio. A haploid, mononucleated genome is preferred because effects of recessive mutations on the phenotype can be recognized without the need of homozygosis. Multiple mutations eventually present in the same cell can be sorted out if procedures for triggering mating under laboratory conditions are available. Beside, breeding offers the valuable opportunity to combine different mutations, contributing to the improvement of the strain for industrial purposes. Since insertion mutagenesis is not targeted, mutants

have to be screened for the desired characteristics. While large-scale screening strategies have been developed, a limiting step is the genetic characterization of insertion sites, that is, the identification of the mutated gene responsible for the observed phenotype. Large genome deletions can occur at the insertion site. The insertion cassette itself can be partially deleted or inserted as concatamer, so that the amplification of genomic regions flanking the insertion through PCR techniques [148] can be difficult. Protocol optimization and improved design of the insertion cassette can decrease these limitations; however, interesting mutants may well resist genetic characterization. Alternative strategies could be performed although time consuming, for instance the *tla1* mutation (see “[Light Use Efficiency](#)” section for further informations about *tla1* mutant) was identified via complementation of the mutant with a library of genomic restriction fragments [60]. In addition, recently developed high throughput sequencing technologies may prove to be effective in rapid re-sequencing of a mutant genome and in localizing the insertion mutation. This strategy is likely to provide future potential for high throughput identification of mutants and annotation of genomes in algae.

### Concluding Remarks

Algae currently used in industry are wild-type strains and mainly cultivated for human/animal food or to extract high-value natural chemicals and pharmaceuticals. However, algae can be also exploited as biomass producers and a source for renewable, ecologically sustainable biofuels. In order to make algae suitable for cultivation in mass culture conditions and economically competitive, domestication of natural strains is required. Selection of the desired characteristics and introduction of new traits through mutagenesis and genetic engineering can substantially improve algae cultivation in the future and make it practicable. Moreover, algae can serve as a “cell factory” for the production of recombinant proteins with pharmaceutical applications. Concerns from public opinion and governments about biosafety can be satisfied and counterbalanced by the potential advantages of transgenic algae, also considering that algae offer the advantage to be grown in a confined environment. Therefore, the use of genetically modified algae for the production

of renewable and sustainable biofuels and pharmaceuticals appears as a likely perspective. Genetic engineering is well useful to generate strains optimized for growth in photobioreactors but uncompetitive in the wild if released.

## Bibliography

### Primary Literature

- Hallmann A (2007) Algal transgenics and biotechnology. *Transgenic Plant J* 1:81–98
- Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, Darzins A (2008) Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. *Plant J* 54:621–639
- Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281:237–240
- Spolaore P, Joannis-Cassan C, Duran E, Isambert A (2006) Commercial applications of microalgae. *J Biosci Bioeng* 101:87–96
- McHugh DJ (2003) A guide to the seaweed industry. *FAO Fish Tech Pap* 441:1–105
- Pulz O, Gross W (2004) Valuable products from biotechnology of microalgae. *Appl Microbiol Biotechnol* 65:635–648
- Borowitzka MA (1999) Commercial production of microalgae: ponds, tanks, tubes and fermenters. *J Biotechnol* 70:313–321
- Banerjee A, Sharma R, Chisti Y, Banerjee UC (2002) *Botryococcus braunii*: a renewable source of hydrocarbons and other chemicals. *Crit Rev Biotechnol* 22:245–279
- Chisti Y (2007) Biodiesel from microalgae. *Biotechnol Adv* 25:294–306
- Chisti Y (2008) Biodiesel from microalgae beats bioethanol. *Trends Biotechnol* 26:126–131
- Xu H, Miao XL, Wu QY (2006) High quality biodiesel production from a microalga *Chlorella protothecoides* by heterotrophic growth in fermenters. *J Biotechnol* 126:499–507
- Ghirardi ML, Zhang L, Lee JW, Flynn T, Seibert M, Greenbaum E, Melis A (2000) Microalgae: a green source of renewable H<sub>2</sub>. *TIBTECH* 18:506–511
- Melis A, Zhang LP, Forestier M, Ghirardi ML, Seibert M (2000) Sustained photobiological hydrogen gas production upon reversible inactivation of oxygen evolution in the green alga *Chlamydomonas reinhardtii*. *Plant Physiol* 122:127–135
- Melis A, Happe T (2004) Trails of green alga hydrogen research - from Hans Gaffron to new frontiers. *Photosynth Res* 80:401–409
- Zhang LP, Melis A (2002) Probing green algal hydrogen production. *Philos Trans R Soc Lond B Biol Sci* 357:1499–1507
- Kruse O, Rupprecht J, Bader KP, Thomas-Hall S, Schenk PM, Finazzi G, Hankamer B (2005) Improved photobiological H<sub>2</sub> production in engineered green algal cells. *J Biol Chem* 280:34170–34177
- Sabine CL, Feely RA, Gruber N, Key RM, Lee K, Bullister JL, Wanninkhof R, Wong CS, Wallace DWR, Tilbrook B, Millero FJ,

- Peng TH, Kozyr A, Ono T, Rios AF (2004) The oceanic sink for anthropogenic CO<sub>2</sub>. *Science* 305:367–371
18. Cox PM, Betts RA, Jones CD, Spall SA, Totterdell IJ (2000) Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* 408:184–187
  19. Caldeira K, Jain AK, Hoffert MI (2003) Climate sensitivity uncertainty and the need for energy without CO<sub>2</sub> emission. *Science* 299:2052–2054
  20. de Moraes MG, Costa JAV (2007) Carbon dioxide fixation by *Chlorella kessleri*, *C. vulgaris*, *Scenedesmus obliquus* and *Spirulina* sp. cultivated in flasks and vertical tubular photobioreactors. *Biotechnol Lett* 29:1349–1352
  21. Keffer JE, Kleinheinz GT (2002) Use of *Chlorella vulgaris* for CO<sub>2</sub> mitigation in a photobioreactor. *J Ind Microbiol Biotechnol* 29:275–280
  22. Otsuki T (2001) A study for the biological CO<sub>2</sub> fixation and utilization system. *Sci Total Environ* 277:21–25
  23. Durrett TP, Benning C, Ohlrogge J (2008) Plant triacylglycerols as feedstocks for the production of biofuels. *Plant J* 54:593–607
  24. Benson AA, Calvin M (1950) Carbon dioxide fixation by green plants. *Annu Rev Plant Physiol* 1:25–42
  25. Melis A (2009) Solar energy conversion efficiencies in photosynthesis: minimizing the chlorophyll antennae to maximize efficiency. *Plant Sci* 177:272–280
  26. Pulz O (2001) Photobioreactors: production systems for phototrophic microorganisms. *Appl Microbiol Biotechnol* 57:287–293
  27. Lee YK (1997) Commercial production of microalgae in the Asia-Pacific rim. *J Appl Phycol* 9:403–411
  28. Borowitzka MA (2005) The mass culture of *Dunaliella salina*. <http://www.fao.org/docrep/field/003/AB728E/AB728E06.htm>
  29. Huesemann MH, Hausmann TS, Bartha R, Aksoy M, Weissman JC, Benemann JR (2008) Biomass productivities in wild type and pigment mutant of *Cyclotella* sp. (Diatom). *Appl Biochem Biotechnol*. doi:10.1007/s12010-008-8298-9
  30. Ugwu CU, Aoyagi H, Uchiyama H (2008) Photobioreactors for mass cultivation of algae. *Bioresour Technol* 99:4021–4028
  31. Mann JE, Myers J (1968) On pigments, growth and photosynthesis of *Phaeodactylum tricornutum*. *J Phycol* 4:349–355
  32. Grima EM, Fernández FGA, Camacho FG, Rubio FC, Chisti Y (2000) Scale-up of tubular photobioreactors. *J Appl Phycol* 12:355–368
  33. Melis A, Neidhardt J, Benemann JR (1998) *Dunaliella salina* (Chlorophyta) with small chlorophyll antenna sizes exhibit higher photosynthetic productivities and photon use efficiencies than normally pigmented cells. *J Appl Phycol* 10:515–525
  34. Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* 127:1309–1321
  35. Bai YL, Lindhout P (2007) Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Ann Bot* 100:1085–1094
  36. Kirk JTO (1994) Light and photosynthesis in aquatic ecosystems, 2nd edn. Cambridge University Press, Cambridge, UK
  37. Holt NE, Fleming GR, Niyogi KK (2004) Toward an understanding of the mechanism of nonphotochemical quenching in green plants. *Biochemistry* 43:8281–8289
  38. Niyogi KK (1999) Photoprotection revisited: genetic and molecular approaches. *Annu Rev Plant Physiol Plant Mol Biol* 50:333–359
  39. Durnford DG, Price JA, Mckim SM, Sarchfield ML (2003) Light-harvesting complex gene expression is controlled by both transcriptional and post-transcriptional mechanisms during photoacclimation in *Chlamydomonas reinhardtii*. *Physiol Plant* 118:193–205
  40. Teramoto H, Nakamori A, Minagawa J, Ono T (2002) Light-intensity-dependent expression of Lhc gene family encoding light-harvesting chlorophyll-a/b proteins of photosystem II in *Chlamydomonas reinhardtii*. *Plant Physiol* 130:325–333
  41. Wobbe L, Schwarz C, Nickelsen J, Kruse O (2008) Translational control of photosynthetic gene expression in phototrophic eukaryotes. *Physiol Plant* 133:507–515
  42. Powles S (1984) Photoinhibition of photosynthesis induced by visible light. *Annu Rev Plant Physiol* 35:15–44
  43. Melis A (1999) Photosystem-II damage and repair cycle in chloroplasts: what modulates the rate of photodamage in vivo? *Trends Plant Sci* 4:130–135
  44. Nakajima Y, Tsuzuki M, Ueda R (1998) Reduced photoinhibition of a phycocyanin deficient mutant of *Synechocystis* PCC 6714. *J Appl Phycol* 10:447–452
  45. Naus J, Melis A (1991) Changes of photosystem stoichiometry during cell growth in *Dunaliella salina* cultures. *Plant Cell Physiol* 32:569–575
  46. Neidhardt J, Benemann JR, Zhang L, Melis A (1998) Photosystem-II repair and chloroplast recovery from irradiance stress: relationship between chronic photoinhibition, light-harvesting chlorophyll antenna size and photosynthetic productivity in *Dunaliella salina* (green algae). *Photosynth Res* 56:175–184
  47. Vener AV, Vankan PJM, Rich PR, Ohad I, Andersson B (1997) Plastoquinol at the quinol oxidation site of reduced cytochrome bf mediates signal transduction between light and protein phosphorylation: thylakoid protein kinase deactivation by a single-turnover flash. *Proc Natl Acad Sci USA* 94:1585–1590
  48. Kok B (1953) Experiments on photosynthesis by *Chlorella* in flashing light. In: Burlew JS (ed) *Algal culture: from laboratory to pilot plant*. The Carnegie Institution, Washington, DC, pp 63–75
  49. Polle JE, Benemann JR, Tanaka A, Melis A (2000) Photosynthetic apparatus organization and function in the wild type and a chlorophyll b-less mutant of *Chlamydomonas reinhardtii*. Dependence on carbon source. *Planta* 211:335–344
  50. Tanaka A, Ito H, Tanaka R, Tanaka NK, Yoshida K, Okada K (1998) Chlorophyll a oxygenase (CAO) is involved in chlorophyll b formation from chlorophyll a. *Proc Natl Acad Sci USA* 95:12719–12723
  51. Polle JE, Niyogi KK, Melis A (2001) Absence of lutein, violaxanthin and neoxanthin affects the functional chlorophyll

- antenna size of photosystem-II but not that of photosystem-I in the green alga *Chlamydomonas reinhardtii*. *Plant Cell Physiol* 42:482–491
52. Elrad D, Niyogi KK, Grossman AR (2002) A major light-harvesting polypeptide of photosystem II functions in thermal dissipation. *Plant Cell* 14:1801–1816
  53. Baroli I, Gutman BL, Ledford HK, Shin JW, Chin BL, Havaux M, Niyogi KK (2004) Photo-oxidative stress in a xanthophyll-deficient mutant of *Chlamydomonas*. *J Biol Chem* 279: 6337–6344
  54. Dall'osto L, Cazzaniga S, Havaux M, Bassi R (2010) Enhanced photoprotection by protein-bound vs free xanthophyll pools: a comparative analysis of chlorophyll b and xanthophyll biosynthesis mutants. *Mol Plant* 3:576–593
  55. Niyogi KK, Bjorkman O, Grossman AR (1997) The roles of specific xanthophylls in photoprotection. *Proc Natl Acad Sci USA* 94:14162–14167
  56. Polle JEW, Kanakagiri SD, Melis A (2003) *tla1*, a DNA insertional transformant of the green alga *Chlamydomonas reinhardtii* with a truncated light-harvesting chlorophyll antenna size. *Planta* 217:49–59
  57. Mussgnug JH, Wobbe L, Elles I, Claus C, Hamilton M, Fink A, Kahmann U, Kapazoglou A, Mullineaux CW, Hippler M, Nickelsen J, Nixon PJ, Kruse O (2005) NAB1 is an RNA binding protein involved in the light-regulated differential expression of the light-harvesting antenna of *Chlamydomonas reinhardtii*. *Plant Cell* 17:3409–3421
  58. Beckmann J, Lehr F, Finazzi G, Hankamer B, Posten C, Wobbe L, Kruse O (2009) Improvement of light to biomass conversion by de-regulation of light-harvesting protein translation in *Chlamydomonas reinhardtii*. *J Biotechnol* 142: 70–77
  59. Mussgnug JH, Thomas-Hall S, Rupprecht J, Foo A, Klassen V, McDowall A, Schenk PM, Kruse O, Hankamer B (2007) Engineering photosynthetic light capture: impacts on improved solar energy to biomass conversion. *Plant Biotechnol J* 5:802–814
  60. Tetali SD, Mitra M, Melis A (2007) Development of the light-harvesting chlorophyll antenna in the green alga *Chlamydomonas reinhardtii* is regulated by the novel *Tla1* gene. *Planta* 225:813–829
  61. Glick RE, Melis A (1988) Minimum photosynthetic unit size in system-I and system-II of barley chloroplasts. *Biochim Biophys Acta* 934:151–155
  62. Falkowski PG, Owens TG (1980) Light-shade adaptation - 2 strategies in marine-phytoplankton. *Plant Physiol* 66:592–595
  63. Dall'osto L, Caffarri S, Bassi R (2005) A mechanism of nonphotochemical energy dissipation, independent from PsbS, revealed by a conformational change in the antenna protein CP26. *Plant Cell* 17:1217–1232
  64. de Bianchi S, Ballottari M, Dall'osto L, Bassi R (2010) Regulation of plant light harvesting by thermal dissipation of excess energy. *Biochem Soc Trans* 38:651–660
  65. Li XP, Bjorkman O, Shih C, Grossman AR, Rosenquist M, Jansson S, Niyogi KK (2000) A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* 403:391–395
  66. Li XP, Gilmore AM, Caffarri S, Bassi R, Golan T, Kramer D, Niyogi KK (2004) Regulation of photosynthetic light harvesting involves intrathylakoid lumen pH sensing by the PsbS protein. *J Biol Chem* 279:22866–22874
  67. Bonente G, Howes BD, Caffarri S, Smulevich G, Bassi R (2008) Interactions between the photosystem II subunit PsbS and xanthophylls studied in vivo and in vitro. *J Biol Chem* 283:8434–8445
  68. Ahn TK, Avenson TJ, Ballottari M, Cheng YC, Niyogi KK, Bassi R, Fleming GR (2008) Architecture of a charge-transfer state regulating light harvesting in a plant antenna protein. *Science* 320:794–797
  69. Holt NE, Zigmantas D, Valkunas L, Li XP, Niyogi KK, Fleming GR (2005) Carotenoid cation formation and the regulation of photosynthetic light harvesting. *Science* 307:433–436
  70. Bonente G, Passarini F, Cazzaniga S, Mancone C, Buia MC, Tripodi M, Bassi R, Caffarri S (2008) The occurrence of the *psbS* gene product in *Chlamydomonas reinhardtii* and in other photosynthetic organisms and its correlation with energy quenching. *Photochem Photobiol* 84: 1359–1370
  71. Niyogi KK, Bjorkman O, Grossman AR (1997) *Chlamydomonas* xanthophyll cycle mutants identified by video imaging of chlorophyll fluorescence quenching. *Plant Cell* 9:1369–1380
  72. Alboresi A, Gerotto C, Giacometti GM, Bassi R, Morosinotto T (2010) *Physcomitrella patens* mutants affected on heat dissipation clarify the evolution of photoprotection mechanisms upon land colonization. *Proc Natl Acad Sci USA* 107: 11128–11133
  73. Peers G, Truong TB, Ostendorf E, Busch A, Elrad D, Grossman AR, Hippler M, Niyogi KK (2009) An ancient light-harvesting protein is critical for the regulation of algal photosynthesis. *Nature* 462:518–U215
  74. Wilson A, Ajlani G, Verbavatz JM, Vass I, Kerfeld CA, Kirilovsky D (2006) A soluble carotenoid protein involved in phycobilisome-related energy dissipation in cyanobacteria. *Plant Cell* 18:992–1007
  75. Eberhard S, Finazzi G, Wollman FA (2008) The dynamics of photosynthesis. *Annu Rev Genet* 42:463–515
  76. Kramer DM, Avenson TJ, Edwards GE (2004) Dynamic flexibility in the light reactions of photosynthesis governed by both electron and proton reactions. *Trends Plant Sci* 9:349–357
  77. Hoefnagel MHN, Atkin OK, Wiskich JT (1998) Interdependence between chloroplasts and mitochondria in the light and the dark. *Biochim Biophys Acta: Bioenerg* 1366:235–255
  78. Raghavendra AS, Padmasree K (2003) Beneficial interactions of mitochondrial metabolism with photosynthetic carbon assimilation. *Trends Plant Sci* 8:546–553
  79. Scheibe R (1987) NADP-malate dehydrogenase in C3-plants: regulation and role of a light-activated enzyme. *Physiol Plant* 71:393–400

80. Cardol P, Gloire G, Havaux M, Remacle C, Matagne R, Franck F (2003) Photosynthesis and state transitions in mitochondrial mutants of *Chlamydomonas reinhardtii* affected in respiration. *Plant Physiol* 133:2010–2020
81. Cardol P, Alric J, Girard-Bascou J, Franck F, Wollman FA, Finazzi G (2009) Impaired respiration discloses the physiological significance of state transitions in *Chlamydomonas*. *Proc Natl Acad Sci USA* 106:15979–15984
82. Peltier G, Cournac L (2002) Chlororespiration. *Annu Rev Plant Biol* 53:523–550
83. Scherer S (1990) Do photosynthetic and respiratory electron-transport chains share redox proteins. *Trends Biochem Sci* 15:458–462
84. Bennoun P (1982) Evidence for a respiratory chain in the chloroplast. *Proc Natl Acad Sci USA* 79:4352–4356
85. Jans F, Mignolet E, Houyoux PA, Cardol P, Ghysels B, Cuine S, Cournac L, Peltier G, Remacle C, Franck F (2008) A type II NAD (P)H dehydrogenase mediates light-independent plastoquinone reduction in the chloroplast of *Chlamydomonas*. *Proc Natl Acad Sci USA* 105:20546–20551
86. Mus F, Cournac L, Cardellini W, Caruana A, Peltier G (2005) Inhibitor studies on non-photochemical plastoquinone reduction and H<sub>2</sub> photoproduction in *Chlamydomonas reinhardtii*. *Biochim Biophys Acta: Bioenerg* 1708:322–332
87. Bailey S, Melis A, Mackey KRM, Cardol P, Finazzi G, van Dijken G, Berg GM, Arrigo K, Shrager J, Grossman A (2008) Alternative photosynthetic electron flow to oxygen in marine *Synechococcus*. *Biochim Biophys Acta: Bioenerg* 1777:269–276
88. Cardol P, Bailleul B, Rappaport F, Derelle E, Beal D, Breyton C, Bailey S, Wollman FA, Grossman A, Moreau H, Finazzi G (2008) An original adaptation of photosynthesis in the marine green alga *Ostreococcus*. *Proc Natl Acad Sci USA* 105:7881–7886
89. Carol P, Stevenson D, Bisanz C, Breitenbach J, Sandmann G, Mache R, Coupland G, Kuntz M (1999) Mutations in the *Arabidopsis* gene *immutans* cause a variegated phenotype by inactivating a chloroplast terminal oxidase associated with phytoene desaturation. *Plant Cell* 11:57–68
90. Cournac L, Redding K, Ravenel J, Rumeau D, Josse EM, Kuntz M, Peltier G (2000) Electron flow between photosystem II and oxygen in chloroplasts of photosystem I-deficient algae is mediated by a quinol oxidase involved in chlororespiration. *J Biol Chem* 275:17256–17262
91. Wu DY, Wright DA, Wetzel C, Voytas DF, Rodermel S (1999) The *immutans* variegation locus of *Arabidopsis* defines a mitochondrial alternative oxidase homolog that functions during early chloroplast biogenesis. *Plant Cell* 11:43–55
92. Finazzi G, Furia A, Barbagallo RP, Forti G (1999) State transitions, cyclic and linear electron transport and photophosphorylation in *Chlamydomonas reinhardtii*. *Biochim Biophys Acta: Bioenerg* 1413:117–129
93. Finazzi G, Rappaport F, Furia A, Fleischmann M, Rochaix JD, Zito F, Forti G (2002) Involvement of state transitions in the switch between linear and cyclic electron flow in *Chlamydomonas reinhardtii*. *EMBO Rep* 3:280–285
94. Vallon O, Bulte L, Dainese P, Olive J, Bassi R, Wollman FA (1991) Lateral redistribution of cytochrome B6/F complexes along thylakoid membranes upon state transitions. *Proc Natl Acad Sci USA* 88:8262–8266
95. Liska AJ, Shevchenko A, Pick U, Katz A (2004) Enhanced photosynthesis and redox energy production contribute to salinity tolerance in *Dunaliella* as revealed by homology-based proteomics. *Plant Physiol* 136:2806–2817
96. Depege N, Bellaïfiore S, Rochaix JD (2003) Role of chloroplast protein kinase Stt7 in LHClI phosphorylation and state transition in *Chlamydomonas*. *Science* 299:1572–1575
97. Fleischmann MM, Ravanel S, Delosme R, Olive J, Zito F, Wollman FA, Rochaix JD (1999) Isolation and characterization of photoautotrophic mutants of *Chlamydomonas reinhardtii* deficient in state transition. *J Biol Chem* 274:30987–30994
98. Rochaix JD (2007) Role of thylakoid protein kinases in photosynthetic acclimation. *FEBS Lett* 581:2768–2775
99. Bonaventura C, Myers J (1969) Fluorescence and oxygen evolution from *Chlorella pyrenoidosa*. *Biochim Biophys Acta* 189:366–383
100. Murata N (1969) Control of excitation transfer in photosynthesis. I. Light-induced change of chlorophyll a fluorescence in *Porphyridium cruentum*. *Biochim Biophys Acta* 172:242–251
101. Delosme R, Olive J, Wollman FA (1996) Changes in light energy distribution upon state transitions: an in vivo photoacoustic study of the wild type and photosynthesis mutants from *Chlamydomonas reinhardtii*. *Biochim Biophys Acta: Bioenerg* 1273:150–158
102. Bellaïfiore S, Bameche F, Peltier G, Rochaix JD (2005) State transitions and light adaptation require chloroplast thylakoid protein kinase STN7. *Nature* 433:892–895
103. Ben Amotz A, Shaish A, Avron M (1989) Mode of action of the massively accumulated beta-carotene of *Dunaliella bardawil* in protecting the alga against damage by excess irradiation. *Plant Physiol* 91:1040–1043
104. Metzger P, Largeau C (2005) *Botryococcus braunii*: a rich source for hydrocarbons and related ether lipids. *Appl Microbiol Biotechnol* 66:486–496
105. Rabbani S, Beyer P, Von Lintig J, Hugueney P, Kleinig H (1998) Induced beta-carotene synthesis driven by triacylglycerol deposition in the unicellular alga *Dunaliella bardawil*. *Plant Physiol* 116:1239–1248
106. Zhekisheva M, Boussiba S, Khozin-Goldberg I, Zarka A, Cohen Z (2002) Accumulation of oleic acid in *Haematococcus pluvialis* (Chlorophyceae) under nitrogen starvation or high light is correlated with that of astaxanthin esters. *J Phycol* 38:325–331
107. Muradyan EA, Klyachko-Gurvich GL, Tsoglin LN, Sergeyenkov TV, Pronina NA (2004) Changes in lipid metabolism during adaptation of the *Dunaliella salina* photosynthetic apparatus to high CO<sub>2</sub> concentration. *Russ J Plant Physiol* 51:53–62



108. Rodolfi L, Zittelli GC, Bassi N, Padovani G, Biondi N, Bonini G, Tredici MR (2009) Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. *Biotechnol Bioeng* 102:100–112
109. Mus F, Dubini A, Seibert M, Posewitz MC, Grossman AR (2007) Anaerobic acclimation in *Chlamydomonas reinhardtii* - anoxic gene expression, hydrogenase induction, and metabolic pathways. *J Biol Chem* 282:25475–25486
110. Roessler PG (1990) Purification and characterization of acetyl-CoA carboxylase from the diatom *Cyclotella cryptica*. *Plant Physiol* 92:73–78
111. Roessler PG, Ohlrogge JB (1993) Cloning and characterization of the gene that encodes acetyl-Coenzyme-A carboxylase in the alga *Cyclotella cryptica*. *J Biol Chem* 268:19254–19259
112. Roessler PG, Bleibaum JL, Thompson GA, Ohlrogge JB (1994) Characteristics of the gene that encodes acetyl-CoA carboxylase in the diatom *Cyclotella cryptica*. *Recombinant DNA Technol* 11:219–225
113. Sheehan J, Dunahay T, Benemann J, Roessler PG (1998) US Department of Energy's Office of Fuels Development. A look back at the US Department of Energy's Aquatic Species Program – Biodiesel from algae. Close out report TP-580-24190. National Renewable Energy Laboratory, Golden
114. Shintani DK, Ohlrogge JB (1995) Feedback inhibition of fatty-acid synthesis in tobacco suspension cells. *Plant J* 7:577–587
115. Klaus D, Ohlrogge JB, Neuhaus HE, Dormann P (2004) Increased fatty acid production in potato by engineering of acetyl-CoA carboxylase. *Planta* 219:389–396
116. Roessler K, Shintani D, Savage L, Boddupalli S, Ohlrogge J (1997) Targeting of the Arabidopsis homomeric acetyl-coenzyme A carboxylase to plastids of rapeseeds. *Plant Physiol* 113:75–81
117. Li Y, Han D, Hu G, Dauvillee D, Sommerfeld M, Ball S, Hu Q (2010) *Chlamydomonas* starchless mutant defective in ADP-glucose pyrophosphorylase hyper-accumulates triacylglycerol. *Metab Eng* 12(4):387–391
118. Arisz SA, van Himbergen JAJ, Musgrave A, van den Ende H, Munnik T (2000) Polar glycerolipids of *Chlamydomonas moewusii*. *Phytochemistry* 53:265–270
119. Guschina IA, Harwood JL (2006) Lipids and lipid metabolism in eukaryotic algae. *Prog Lipid Res* 45:160–186
120. Poerschmann J, Spijkerman E, Langer U (2004) Fatty acid patterns in *Chlamydomonas* sp. as a marker for nutritional regimes and temperature under extremely acidic conditions. *Microb Ecol* 48:78–89
121. Kimura K, Yamaoka M, Kamisaka Y (2004) Rapid estimation of lipids in oleaginous fungi and yeasts using Nile red fluorescence. *J Microbiol Meth* 56:331–338
122. Grobbelaar JU (2004) Algal nutrition. In: Richmond A (ed) *Handbook of microalgal culture: biotechnology and applied phycology*. Blackwell, Oxford, pp 97–115
123. Pyle DJ, Garcia RA, Wen ZY (2008) Producing docosahexaenoic acid (DHA)-rich algae from biodiesel-derived crude glycerol: effects of impurities on DHA production and algal biomass composition. *J Agric Food Chem* 56:3933–3939
124. Geng D, Wang Y, Wang P, Li W, Sun Y (2003) Stable expression of hepatitis B surface antigen gene in *Dunaliella salina* (Chlorophyta). *J Appl Phycol* 15:451–456
125. Mayfield SP, Franklin SE, Lerner RA (2003) Expression and assembly of a fully active antibody in algae. *Proc Natl Acad Sci USA* 100:438–442
126. Mayfield SP, Franklin SE (2005) Expression of human antibodies in eukaryotic micro-algae. *Vaccine* 23:1828–1832
127. Sayre RT, Wagner RE, Sirporanadulsil S, Farias C (2001) Transgenic algae for delivery antigens to animals. *Int. Patent Number W.O.01/98335 A2*
128. Sun M, Qian K, Su N, Chang H, Liu J, Shen G (2003) Foot-and-mouth disease virus VP1 protein fused with cholera toxin B subunit expressed in *Chlamydomonas reinhardtii* chloroplast. *Biotechnol Lett* 25:1087–1092
129. Walker TL, Purton S, Becker DK, Collet C (2005) Microalgae as bioreactors. *Plant Cell Rep* 24:629–641
130. Govoni C, Morosinotto T, Giuliano G, Bassi R (2007) Exploiting photosynthesis for biofuel production. In: Pavesi L, Fauchet (eds) *Biophotonics*. Springer-Verlag, Berlin Heidelberg, pp 15–28
131. Gaffron H, Rubin J (1942) Fermentative and photochemical production of hydrogen in algae. *J Gen Physiol* 26:219–240
132. Surzycki R, Cournac L, Peltier G, Rochaix JD (2007) Potential for hydrogen production with inducible chloroplast gene expression in *Chlamydomonas*. *Proc Natl Acad Sci USA* 104:17548–17553
133. Purton S (2007) Tools and techniques for chloroplast transformation of *Chlamydomonas*. *Adv Exp Med Biol* 616:34–45
134. Molina GE, Belarbi EH, Acien Fernandez FG, Robles MA, Chisti Y (2003) Recovery of microalgal biomass and metabolites: process options and economics. *Biotechnol Adv* 20:491–515
135. Belarbi EH, Molina E, Chisti Y (2000) A process for high yield and scaleable recovery of high purity eicosapentaenoic acid esters from microalgae and fish oil. *Enzyme Microb Technol* 26:516–529
136. Stevens DR, Rochaix JD, Purton S (1996) The bacterial phleomycin resistance gene ble as a dominant selectable marker in *Chlamydomonas*. *Mol Gen Genet* 251:23–30
137. Dunahay TG, Jarvis EE, Zeiler KG, Roessler PG, Brown LM (1992) Genetic-engineering of microalgae for fuel production - scientific note. *Appl Biochem Biotechnol* 34–5:331–339
138. Falcitatore A, Casotti R, Leblanc C, Abrescia C, Bowler C (1999) Transformation of nonselectable reporter genes in marine diatoms. *Mar Biotechnol* 1:239–251
139. Merendino L, Perron K, Rahire M, Howald I, Rochaix JD, Goldschmidt-Clermont M (2006) A novel multifunctional factor involved in trans-splicing of chloroplast introns in *Chlamydomonas*. *Nucleic Acids Res* 34:262–274
140. Sizova I, Fuhrmann M, Hegemann P (2001) A *Streptomyces rimosus* aphVIII gene coding for a new type phosphotransferase

provides stable antibiotic resistance to *Chlamydomonas reinhardtii*. *Gene* 277:221–229

141. Fischer N, Rochaix JD (2001) The flanking regions of *PsaD* drive efficient gene expression in the nucleus of the green alga *Chlamydomonas reinhardtii*. *Mol Genet Genomics* 265:888–894
142. Schroda M, Blocker D, Beck CF (2000) The HSP70A promoter as a tool for the improved expression of transgenes in *Chlamydomonas*. *Plant J* 21:121–131
143. Hallmann A, Sumper M (1994) An inducible arylsulfatase of *Volvox carteri* with properties suitable for a reporter-gene system - purification, characterization and molecular-cloning. *Eur J Biochem* 221:143–150
144. Lumbreras V, Stevens DR, Purton S (1998) Efficient foreign gene expression in *Chlamydomonas reinhardtii* mediated by an endogenous intron. *Plant J* 14:441–447
145. Cerutti H, Johnson AM, Gillham NW, Boynton JE (1997) Epigenetic silencing of a foreign gene in nuclear transformants of *Chlamydomonas*. *Plant Cell* 9:925–945
146. Mayfield SP, Manuell AL, Chen S, Wu J, Tran M, Siefker D, Muto M, Marin-Navarro J (2007) *Chlamydomonas reinhardtii* chloroplasts as protein factories. *Curr Opin Biotechnol* 18:126–133
147. McCallum CM, Comai L, Greene EA, Henikoff S (2000) Targeted screening for induced mutations. *Nat Biotechnol* 18:455–457
148. Dent RM, Haglund CM, Chin BL, Kobayashi MC, Niyogi KK (2005) Functional genomics of eukaryotic photosynthesis using insertional mutagenesis of *Chlamydomonas reinhardtii*. *Plant Physiol* 137:545–556

## Books and Reviews

- Beer LL, Boyd ES, Peters JW, Posewitz MC (2009) Engineering algae for biohydrogen and biofuel production. *Curr Opin Biotechnol* 20:264–271
- Dolley TP, Moyle PR (2003) History and overview of the U.S. diatomite mining industry, with emphasis on the Western United States. *Contrib Ind Min Res: US Geol Surv Bull* 2209: E1–E8
- Greenwell HC, Laurens LML, Shields RJ, Lovitt RW, Flynn KJ (2010) Placing microalgae on the biofuels priority list: a review of the technological challenges. *J R Soc Interface* 7:703–726
- Rupprecht J (2009) From systems biology to fuel - *Chlamydomonas reinhardtii* as a model for a systems biology approach to improve biohydrogen production. *J Biotechnol* 142:10–20
- Scott SA, Davey MP, Dennis JS, Horst I, Howe CJ, Lea-Smith DJ, Smith AG (2010) Biodiesel from algae: challenges and prospects. *Curr Opin Biotechnol* 21:277–286
- Tredici MR (2003) Mass production of microalgae photobioreactors. In: Richmond A (ed) *Handbook of microalgal culture*. Blackwell Publishing, Malden, pp 178–214
- Van den Hoek C, Mann DG, Jahns HM (1995) *Algae: an introduction to phycology*. Cambridge University Press, Cambridge, UK

## Alkaline Membrane Fuel Cells

ROBERT C. T. SLADE, JAMIE P. KIZEWSKI, SIMON D. POYNTON, RONG ZENG, JOHN R. VARCOE  
Department of Chemistry, University of Surrey,  
Guildford, UK

### Article Outline

Glossary  
Definition  
Introduction  
An Overview of Alkaline Anion-Exchange Membranes (AEMs)  
Properties of AEMs  
Synthetic Routes to AEMs  
Development of Alkaline Ionomers  
AMFCs with H<sub>2</sub> as Fuel  
The Application of AEMs in Other Fuel Cell Types  
A Brief Overview of the Patent Literature  
Future Directions and Research Challenges  
Acknowledgment  
Bibliography

### Glossary

**AEM** Alkaline (anion) exchange membrane.  
**AFC** Alkaline fuel cell  
**AMFC** Alkaline membrane fuel cell (also known as APEMFC)  
**DMFC** Direct methanol fuel cell  
**MEA** Membrane electrode assembly  
**OCV** Open circuit voltage  
**PEM** Proton-exchange membrane  
**PEMFC** Proton-exchange membrane fuel cell  
**QA** Quaternary ammonium  
**RG-AEM** Radiation-grafted alkaline (anion) exchange membrane

### Definition

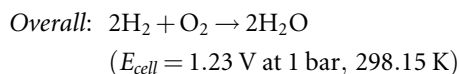
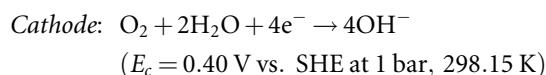
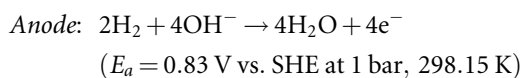
The disruptive approach of applying alkaline anion-exchange membranes (AEMs) in alkaline membrane fuel cells (AMFCs) potentially meets several of the challenges facing other approaches to low temperature fuel cells, including the otherwise high catalyst and fuel costs. Thus, the move to alkaline conditions at the electrodes

opens the potential use of a range of low cost non-precious-metal catalysts, as opposed to the otherwise necessary use of platinum-group-metal (PGM) based catalysts. Further, it becomes possible to consider hydrogen fuels containing substantial amounts of impurities, whereas an acidic membrane approach (that in proton exchange membrane fuel cells, PEMFCs) requires high-purity gases and PGM catalysts.

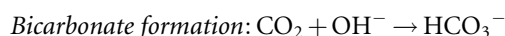
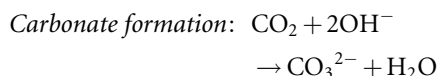
## Introduction

The first entry in the AMFC area was published in 2005 [1], since when activity and interest have continued to increase steeply internationally. Zeng and Varcoe have recently reviewed the developing patent literature [2]. Some researchers have recently termed these systems HEMFCs, hydroxide exchange membrane fuel cells; that terminology is not fully appropriate in view of the complex hydroxide/hydrogen carbonate/carbonate equilibria that are present (even after handling membranes in air), inevitable in the use of air (containing CO<sub>2</sub>(g)) as source of oxidant and also produced in the oxidation of methanol in direct methanol fuel cells (DMFCs). Another acronym applied to these systems, APEMFCs, is also potentially confusing, the first three words of the full form then seemingly being “alkaline proton exchange” (following the very similar PEMFC acronym) but being readable alternatively as alkaline polymer electrolyte membrane fuel cells (possibly AEMFC would be clearer).

Alkaline fuel cells (AFCs, hydrogen-fuelled cells with an alkaline liquid electrolyte such as KOH(aq)) are the best performing of all known conventional hydrogen–oxygen fuel cells operable at temperatures below 200°C. This is due to the facile kinetics at the cathode and at the anode; cheaper non-noble metal catalysts can be used (such as nickel and silver [3, 4]), reducing cost. McLean et al. gave comprehensive review of alkaline fuel cell technology [5]. The associated fuel cell reactions both for a traditional AFC and also for an AMFC are:



A major issue with traditional AFCs is that of electrolyte and electrode degradation caused by the formation of carbonate/bicarbonate (CO<sub>3</sub><sup>2-</sup>/HCO<sub>3</sub><sup>-</sup>) on reaction of OH<sup>-</sup> ions with CO<sub>2</sub> contamination in the oxidant gas stream [5–7]:



The major cause of the degrading performance of AFCs is the consequent precipitation of metal carbonate crystals (most commonly Na<sub>2</sub>CO<sub>3</sub> or K<sub>2</sub>CO<sub>3</sub>, depending on the alkaline electrolyte used) in the electrolyte-filled pores of the electrodes, blocking pores and mechanically disrupting and destroying active layers.

The cost of fuel cells still retards commercialization in most markets. AFCs are promising on a cost basis mainly because cheap and relatively abundant non-platinum-group metals (non-PGM) are viable catalysts, but are hindered by degradation due to formation of precipitates as above. Catalyst electrokinetics (for fuel oxidation and oxygen reduction) are improved in alkaline, as opposed to acidic, conditions (the acid-stability criterion precludes the use of most non-PGM catalysts in PEMFCs). The replacement of the KOH (aq) electrolyte with an alkaline electrolyte membrane (AEM), to give AMFCs, retains the electrocatalytic advantages but introduces CO<sub>2</sub> tolerance (there being no mobile cations that could give carbonate/bicarbonate precipitates) with the additional advantage of being an all-solid-state fuel cell (as with PEMFCs – i.e., no seeping out of KOH(aq)). Additionally, thin (low electronic resistance) and easily stamped (cheap) metal mono/bipolar plates can be used, with reduced corrosion-derived problems at high pH (the cost of bipolar plates for PEMFCs can be as much as one third of the cost of the stacks themselves). A key and yet to be convincingly met requirement is, however, the development of a dispersible alkaline ionomer (sometimes termed an anionomer) to maximize ionic contact between the catalyst reaction sites and the ion-conductive membranes. As in the case of AFCs, water

is produced at anode and consumed at cathode in AMFCs (when fuelled with hydrogen and with four electron reduction of oxygen at the cathode), which is fundamentally different to what occurs in PEMFCs containing acidic electrolytes; this can cause high overpotentials at AMFC anodes, due to suspected flooding [8].

The use of an AEM as a solid electrolyte including no metal cations prevents precipitation of carbonate/bicarbonate salts. (The electrolyte containing the cationic groups is already a solid.) The carbonation process is quick even if the AEM has been exposed to the air for only a short time [9, 10]. The conductivities of the AEMs in  $\text{OH}^-$  form may have been underestimated because most studies to date have not disclosed vigorous  $\text{CO}_2$  exclusion procedures during conductivity measurements. It has been hypothesized that  $\text{OH}^-$  ion conductivities in AEMs can be estimated by measuring the ionic conductivities of  $\text{HCO}_3^-$  from AEMs and multiplying by 3.8 [11]. This carbonation process may not to be a serious problem due to an in situ “self-purging mechanism” because  $\text{OH}^-$  anions are continuously generated at the cathode in AMFCs [12].

AEMs are solid polymer electrolyte membranes that contain positive ionic groups (typically quaternary ammonium (QA) functional groups such as  $\text{poly-NMe}_3^+$ ) and mobile negatively charged anions. A widely quoted concern with AEMs is membrane stability, especially at elevated temperatures [13, 14]. The general issues are:

1. The diffusion coefficients and mobilities of  $\text{OH}^-$  anions are less than that of  $\text{H}^+$  in most media, and QA ionic groups are less dissociated than the typical sulfonic acid groups ( $\text{p}K_a$  for sulfonic acid groups are typically  $-1$  but for QA groups the related  $\text{p}K_b$  values are around  $+4$ ); there were concerns that AEMs would not possess adequate intrinsic ionic conductivities for application in fuel cells.
2. The  $\text{OH}^-$  anions are effective nucleophiles which potentially cause degradation via (a) a direct nucleophilic displacement and/or (b) a Hofmann elimination reaction when a  $\beta$ -hydrogen is present; methyl ( $-\text{CH}_3$ ) groups may also be displaced by  $\text{OH}^-$  ions forming tertiary amines and methanol [13, 14]. If the AEMs contain good leaving groups (e.g., QA  $-\text{NMe}_3^+$  groups) then the chemical

stability of the AEMs might have been inadequate for use in fuel cells, particularly at elevated temperatures.

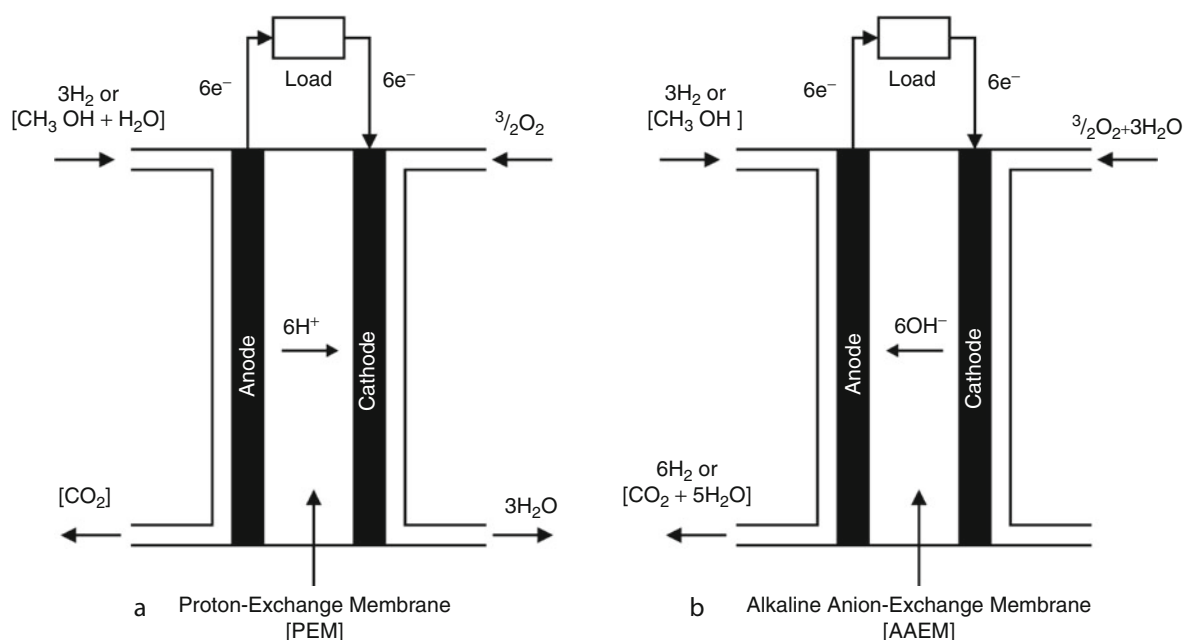
3. Precursor anion-exchange membranes are generally submerged in aqueous  $\text{NaOH/KOH}$  solutions to exchange them to the  $\text{OH}^-$  form AEM; the AEM must have the chemical stability to withstand this process. Despite this, over a decade ago, the stabilities of various benzyltrimethylammonium-based AEMs were found to be stable at up to  $75^\circ\text{C}$  in  $\text{NaOH(aq)}$  at concentrations up to  $6 \text{ mol dm}^{-3}$  for several days [15].

A major potential application of AMFCs is, however, as power sources for at or near room temperature (as for PEMFCs), which means such degradation can be minimal.

This entry considers the current understanding and application of AEMs in hydrogen-fuelled AMFCs and other fuel cell types employing AEMs. Figure 1 enables comparison between hydrogen- or methanol-fuelled PEMFCs and AMFCs; the electrode reactions in a hydrogen-fuelled AMFC are discussed above. In contrast to PEMFCs, operation of an AMFC requires the presence of water as a reagent at the cathode (oxygen reduction reaction, ORR – to form  $\text{OH}^-$ ) and the product water is formed by the hydrogen oxidation reaction (HOR) at the anode (as opposed to being formed at the cathode in PEMFCs). The entry begins with consideration of the main classes of AEMs.

## An Overview of Alkaline Anion-Exchange Membranes (AEMs)

AEMs and alkaline ionomers (anionomers) are key to the successful implementation of AMFCs. Anion-exchange membranes have, for a long time, been used as separation membranes for seawater desalination, the recovery of metal ions from wastewaters, electrodialysis and bio-separation processes, for example [16–26]. These membranes may, however, not be stable or conductive enough to be applied in AMFCs. AEMs used in early AMFC studies were reviewed in 2005 [1] and included polybenzimidazole (PBI) doped with  $\text{KOH}$ , epichlorohydrin polymer quaternized with 1,4-diazabicyclo[2,2,2]octane (DABCO) or quaternized with a 1:1 ratio of DABCO and triethylamine, and commercial membranes such as AHA (Tokuyama Co,



**Alkaline Membrane Fuel Cells. Figure 1**

A schematic presentation of (a) a proton-exchange membrane (PEMFC) and (b) an alkaline membrane fuel cell (AMFC), both fuelled either with  $\text{H}_2$  gas or directly with methanol (DMFC mode). The stoichiometric ratios of reactants and products are shown in each case

Japan), Morgane ADP (Solvey S.A.), Tosflex® SF-17 (Tosoh) and 2259-60 (Pall RAI). Most of these fuel cells containing AEMs were, however, being operated in the presence of aqueous alkaline solutions (containing NaOH or KOH).

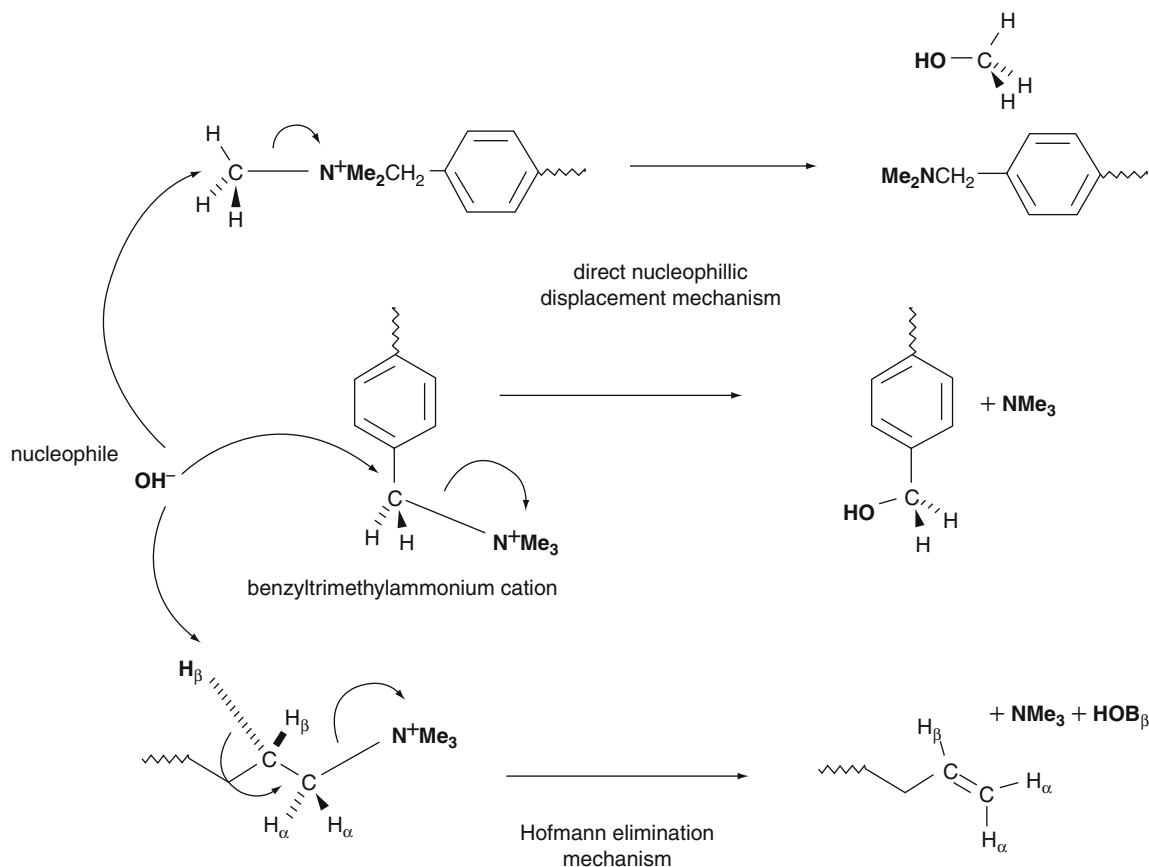
In the UK (at the University of Surrey), several kinds of QA-containing radiation-grafted AEMs (RG-AEMs) based on poly(vinylidene fluoride) PVDF [27–31], poly(tetrafluoroethylene-co-hexafluoropropylene) FEP [27, 28, 32] and poly(ethylene-co-tetrafluoroethylene) ETFE [32, 33], with good ion-exchange capacities (IEC) and ionic conductivities and with sufficient stabilities to test the proof of concept of using AEMs in fuel cells, have been developed. ETFE base films produce the best AEMs for testing in AMFCs. The radiation-grafted methodology (using gamma rays or electron beams for irradiation) allowed for the production of AEMs of different thicknesses, ion-exchange capacity, physical/mechanical properties, and chemistries that facilitated fundamental investigations. A water-insoluble alkaline ionomer that used  $N,N,N',N'$ -tetramethylhexane-1,6-diamine as the joint amination and cross-linking agent first enabled studies

of the performance of metal-cation-free all-solid-state alkaline fuel cells with stable performance into the medium term at  $50^\circ\text{C}$  [33, 34]. Many other groups have now developed new conductive and chemically and thermally stable AEMs (and candidate alkaline ionomers), as discussed below. AEMs in the hydroxide form typically become brittle if allowed to dry; hydroxide forms are typically prepared just before use, from the chloride form analogue.

### Properties of AEMs

The most common class of NaOH-/KOH-free AEMs being investigated for use in fuel cells is based on QA chemistry and has reasonable stability in alkaline environments (especially those AEMs containing benzyltrimethyl ammonium exchange sites). There are three main classes of chemical degradation reaction mechanisms by which nucleophilic  $\text{OH}^-$  anions can remove QA groups. The presence of  $\beta$ -hydrogens allows the Hofmann elimination reaction to occur (Scheme 1), often in parallel to the reactions discussed below, yielding alkene (vinyl) groups; this can give





### Alkaline Membrane Fuel Cells. Scheme 1

Alternative mechanisms for degradation of AEMs by displacement of the trimethylammonium groups by hydroxide anions at elevated temperatures

rise to QA–AEMs that have low thermal and chemical (to alkali) stabilities. If, as in RG–AEMs, no  $\beta$ -hydrogen atoms are present, direct nucleophilic substitution reactions were traditionally thought to take place yielding alcohol and tertiary amine groups. However, recent density functional theory (DFT) calculations and deuterium exchange experiments at Los Alamos National Laboratory (USA) on model small QA-containing compounds indicate that a mechanism involving an ylide intermediate (trimethyl ammonium methylide, also known as a 1,2-dipolar ylide compound) may predominate and be more severe when the AEMs are dehydrated [35].

AEMs from Tokuyama Co have good thermochemical stability. The thin (10  $\mu\text{m}$ ) Tokuyama “fuel cell grade” AEMs (A010, A201 – formerly A006 and

A901) [9, 36–39] and their developmental dispersible alkaline ionomers (A3ver.2 and AS–4) have also been tested at up to 50°C by several research groups in direct alcohol fuel cells [40–42].

Quaternized pyridinium- or phosphonium-based AEMs were thought to have thermochemical stabilities that are not suitable for use in AEM fuel cells but there are reports of recent work on polysulfone-phosphonium-based AEMs and anionomers (see later).

Surrey’s benzyltrimethylammonium-containing S80, S50, and S20 RG–AEMs (the number designating the fully hydrated thicknesses in micrometers) are chemically stable up to 80°C in aqueous KOH [32–34] (aq, 1  $\text{mol dm}^{-3}$ ) and can exhibit ionic conductivities  $> 0.03 \text{ S cm}^{-1}$  at room temperature when fully

hydrated; the ex situ and in situ ionic conductivity of fully hydrated S80 is  $0.06 \text{ S cm}^{-1}$  at  $60^\circ\text{C}$  (cf. Nafion® acidic PEMs are typically  $> 0.1 \text{ S cm}^{-1}$  at these temperatures). AEM conductivities are, however, considerably reduced at humidities  $\text{RH}/\% < 100$  and drop to values between  $0.01$  and  $0.02 \text{ S cm}^{-1}$  after only an hour when exposed to air (especially with very thin membranes) due to the reaction of  $\text{OH}^-$  anions with  $\text{CO}_2$  forming  $\text{CO}_3^{2-}$  and  $\text{HCO}_3^-$  within the membrane. The lower dissociation constant for  $-\text{NMe}_3\text{OH}$  groups (requiring a higher number of water molecules for complete dissociation), compared to  $-\text{SO}_3\text{H}$  groups in PEMs, and a very low number of water molecules directly associated with the ionic groups lead to the poor performance at low RHs (at high humidities much of the water present in AEMs is located in aggregates not directly associated with the ionic groups) [32].

The developing published and patent literature in this area has recently been reviewed by Zeng and Varcoe [2]. Further classes of AEM for fuel cells have included membranes based on quaternized poly(epichlorohydrin), polysulfone, poly(phthalalazinone ether ketone), poly(2,6-dimethyl-1,4-phenylene oxide), and poly(vinyl alcohol) grafted with (2,3-epoxypropyl)trimethylammonium chloride. Common quaternizing agents include alkyl iodides, trialkylamines,  $N,N,N',N'$ -tetramethylalkyl-1, $n$ -diamines, polyethyleneimine, 1,4-diazabicyclo-[2.2.2]-octane (DABCO), and 1-azabicyclo-[2.2.2]-octane. The final two of these have been used extensively; they contain  $\beta$ -hydrogen but their structures do not permit molecular conformations favored in the Hofmann elimination mechanism. Metal-cation-containing AEMs based on doping polymer films (e.g., polybenzimidazole (PBI), poly(vinyl alcohol) and its composites, and biocompatible chitosan) with  $\text{NaOH/KOH(aq)}$  are also being investigated, but the presence of mobile cations may introduce problems associated with precipitation of carbonate salts.

Ionic conductivities of AEMs are generally lower than those of comparable PEMs. This is not surprising as the solution mobility of  $\text{OH}^-$  is one third to one half of that of a  $\text{H}^+$  (depending on the environment and if there  $-\mu_o \times 10^4/\text{cm}^2 \text{ V}^{-1} \text{ s}^{-1} = 20.64$  for  $\text{OH}^-(\text{aq})$  anions and  $36.23$  for  $\text{H}^+(\text{aq})$  cations at  $298 \text{ K}$ ) [1, 43]. One strategy for enhanced ionic conductivities is to increase the ion-exchange capacity (IEC) via synthetic

methodology, but this often leads to a decrease in the mechanical strength due to excessive water uptakes. Another strategy is to synthesize tailored membranes that will exhibit hydrophilic(ionic)–hydrophobic(non-ionic) phase segregation and continuous ionic domains, which is hypothesized to increase ionic conductivities [44, 45].

Recent intensive studies have, however, been reported to lead to AEMs with high ionic conductivities, reportedly comparable to Nafion®. These promising AEMs [11, 45–51] are still to be evaluated in AMFCs. Most hydrocarbon AEMs are soluble in various solvents, which is potentially useful for the formulation of alkaline ionomers required for the preparation of high-performance membrane electrode assemblies (MEAs). If the conductive properties reported can be translated into high power outputs, then AMFC performances comparable to those of PEMFCs can be expected in the near future.

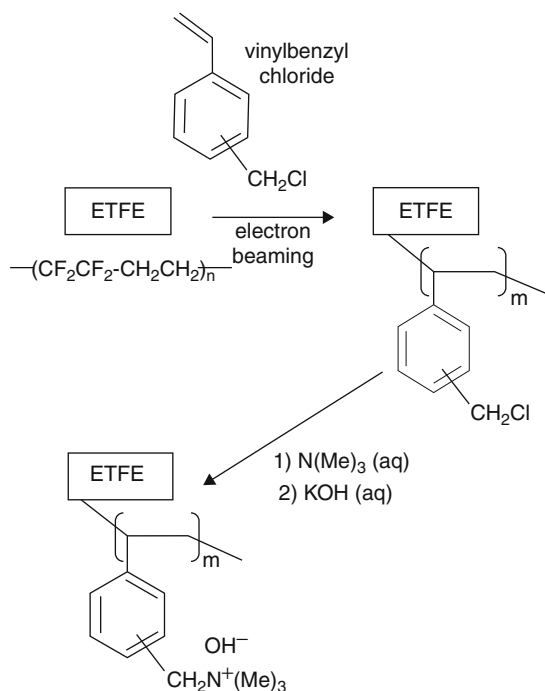
### Synthetic Routes to AEMs

The preparation of applicable AEMs involves a compromise between the properties of the membrane, such as the chemical and thermal stability, ion-exchange capacity (IEC), ion conductivity, mechanical properties, water uptake, and dimensional stability. In general, alkaline anion-exchange polymer electrolytes can be polymerized directly from functionalized monomers, polymerized from monomers with subsequent functionalization or by functionalizing a commercially available polymer. The backbone of the polymer is usually selected for its good chemical and thermal stability and, therefore, typically includes aromatic rings and/or a degree of fluorination: typical polymer classes include polysulfones and polyetherketones, polyimides, poly(phenylene), poly(phthalazinone ether sulfone ketone), polyepichlorohydrin homopolymer, polybenzimidazole (PBI), poly(phenylene oxide), radiation-grafted copolymers, inorganic–organic hybrids, and even perfluorinated membranes such as Nafion. The active functional groups are commonly quaternary ammonium type ( $-\text{NR}_3^+$ ) with a clear preference for trimethylammonium ( $-\text{N}(\text{CH}_3)_3^+$ ) groups ( $\text{pK}_a(\text{H}_2\text{O}) = 9.8$ ).

A suitable AEM will have a high ion-exchange capacity, high ionic conductivity, and thermochemical

stability, but will exhibit a low degree of swelling on hydration. There are several general synthetic methodologies for the preparation of AEMs [52, 53]. Fluorine-containing polymers generally show higher thermal stabilities than hydrocarbon polymers. AEMs based on a poly(arylene ether sulfone) containing fluorine atoms show high ionic conductivities ( $63 \text{ mS cm}^{-1}$  in  $\text{CO}_3^{2-}$  form at  $70^\circ\text{C}$ ) [53]; this is a key result as  $\text{CO}_3^{2-}$  anions have dilute solution mobilities that are less than 33% of that of  $\text{OH}^-$  anions. (It is rare to see  $\text{CO}_3^{2-}$  conductivities above  $30 \text{ mS cm}^{-1}$ .)

Irradiation of polymer films (and powders, etc.) using X-rays,  $\gamma$ -rays, or electron beams (as at Surrey, see previous section) is a flexible way to introduce various functional groups on the polymer backbones (Scheme 2). A wide range of chemically and thermal stable polymers, such as ETFE and FEP, can be chosen as the base films for the production of AEMs. Additionally, there is a wide choice of functional monomers available that can be used to introduce ion-exchange groups into the grafted polymeric chains.



#### Alkaline Membrane Fuel Cells. Scheme 2

The radiation-grafting of vinylbenzyl chloride onto ETFE and subsequent amination and alkali-exchange, yielding alkaline anion-exchange membranes (RG-AEMs)

A common strategy in the synthesis of AEMs is to introduce halogen alkyl groups onto the backbone or side chains of the polymer via chloroalkylation, fluorination, bromination, or chlorination, followed by amination/quaternization and finally ion exchange. Highly carcinogenic chloromethylethers have traditionally been used as the chloroalkylation agent but safer strategies have been introduced, for example, the chloro methylation agent is generated in situ (e.g., [44, 55, 56]). An alternative strategy is the prior introduction of tertiary amine groups into the polymer, followed by quaternization.

The easiest synthetic route is to use inert polymers (as above) doped with concentrated  $\text{KOH}(\text{aq})$  (as above): polybenzimidazole (PBI) [57–61], poly(vinyl alcohol) (PVA) [44], composite polymers such as PVA/hydroxyapatite (PVA/HAP) [62], quaternized-PVA/alumina (QPVA/ $\text{Al}_2\text{O}_3$ ) [63], PVA/titanium oxide (PVA/ $\text{TiO}_2$ ) [64, 65], chitosan and cross-linked chitosan [66–68], copolymers of epichlorohydrin and ethylene oxide [69], and cross-linked PVA/sulfosuccinic acid (10 wt.% SSA) [70] have all been doped with  $\text{KOH}$  and used as AEMs. Patent US5569559 [71] describes the use of polar polymers (most preferred being polyethylene oxide) doped with alkaline metal hydroxides (such as  $\text{KOH}$ ), alkaline earth metal hydroxides or ammonium hydroxides such as tetrabutylammonium hydroxide; PBI doped with  $\text{KOH}$  showed the highest ionic conductivity, comparable to Nafion<sup>®</sup> (a standard acidic proton-exchange membrane, PEM). All of these materials could, however, lead to carbonate precipitates.

Anion-exchange polymers that contain methacrylate, ester, amide, or other carbonyl ( $\text{C}=\text{O}$  double bond) functional groups show low stabilities in alkali as these functional groups are highly reactive to nucleophiles such as  $\text{OH}^-$ .

#### Development of Alkaline Ionomers

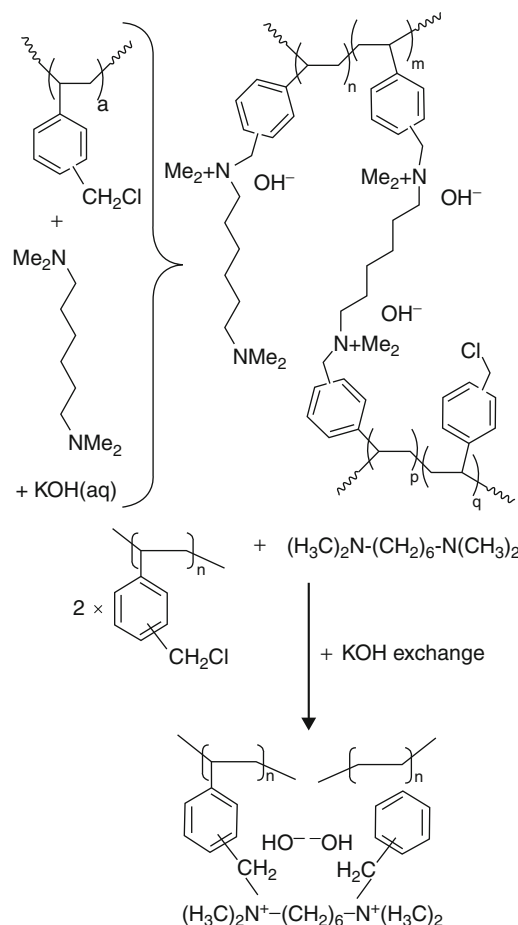
Alkaline analogues to the oligomeric perfluorosulfonic acid dispersions used to produce optimized ionic contact between the catalyst reaction sites and the PEM in PEMFCs have yet to be convincingly developed and this has resulted in the use of nonideal strategies for the fabrication of alkaline MEAs: the use of Nafion dispersions (acidic, cation exchanging); concepts involving

adsorbed potassium hydroxide, and quaternized copolymers made from 4-vinylpyridine monomer at the electrode–AEM interface. Polysulfone-based alkaline ionomers that are compatible with polysulfone AEMs have been developed by various teams. On the commercial front, it has been reported that Fumatech have developed a ionomer concept for use with their commercial Fumasep®FAA AEMs, that uses two precursors that react to form a cross-linked polymer when mixed together, and Tokuyama's alkaline ionomers can be deposited from solution (designated A3 or AS-4) and are chemically compatible with Tokuyama's fuel cell AEMs.

Surrey developed an alkaline ionomer (SION1, Scheme 3) for MEA fabrication, with the objective of developing a system that would allow the testing of different AEMs and electrodes in AMFCs [72]; without this alkaline ionomer, the performances were too low for satisfactory testing. That alkaline ionomer also allowed for the operation of fuel cells in metal-cation-free mode, in contrast to the possible use of doped polymers containing alkali metal hydroxides. An alkaline MEA that was deliberately converted to the  $\text{CO}_3^{2-}$  form operated as well as an  $\text{OH}^-$  form MEA; even more interestingly, the  $\text{CO}_3^{2-}$  content of the carbonated MEA decreased during fuel cell operation with air (containing  $\text{CO}_2$ ) at the cathode. The SION1 ionomer contains  $\beta$ -hydrogen atoms and, therefore, allows the Hofmann elimination degradation mechanism to operate, limiting the thermal stability to below ca.  $60^\circ\text{C}$ . Surrey is currently developing a next-generation  $\beta$ -hydrogen-free alkaline ionomers deposited from aqueous solutions (Patent GB 0814652.4).

Many researchers have reported that alkaline polymers with hydrocarbon backbones can be dissolved in solvents such as DMF, DMAc, and DMSO [47–49, 54, 73, 74], and this allows their use as the ionomer for the preparation of MEAs. Zhuang et al. [73, 75] used a quaternary ammonium polysulfone (QAPS) which can be dissolved in DMF and used as the ionomer; the QAPS polymer (in  $\text{OH}^-$  form:  $\text{IEC} = 1.08 \text{ mmol g}^{-1}$ ) was used to fabricate the membrane, while a QAPS film ( $\text{OH}^-$  form:  $\text{IEC} = 1.18 \text{ meq g}^{-1}$ ) was used as the ionomer.

The development of alkaline ionomers is one of the major challenges for the development of high-performance alkaline membrane containing fuel cells.



### Alkaline Membrane Fuel Cells. Scheme 3

The formation of SION1 alkaline ionomer (anionomer) by cross-linkage of poly(vinylbenzyl chloride) with tetramethylhexanediamine (TMHDA)

In state-of-the-art AMFC technology, the role of the ionomer has not been fully investigated or understood; alkaline ionomer solutions/dispersions, comparable to the Nafion dispersions used in PEMFCs, remain highly sought after.

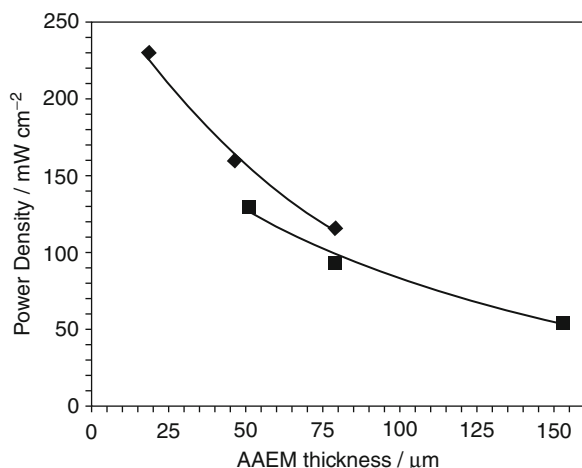
### AMFCs with $\text{H}_2$ as Fuel

Fundamental studies at Surrey revealed the following [31–34]:

*A major source of performance loss is the mass transport of  $\text{H}_2\text{O}$  to the ORR reaction sites. S80/S50/S20 radiation-grafted RG–AEMs, Solvay's Morgane®–ADP, and Tokuyama's A201 were evaluated in  $\text{H}_2/\text{O}_2$*

fuel cells at Surrey using both commercial (prefabricated) and in-house produced electrodes containing SION1 ionomer [72]. As discussed earlier, water is required as a stoichiometric reactant for the ORR at the cathode when using AEMs (unlike in PEMFCs).

The primary source of  $H_2O$  to the cathode appears to be the back transport of the  $H_2O$  electro-generated at the anode. Even with the use of fully hydrated gas supplies, evidence for this was that enhanced back transport of  $H_2O$  from the anode, and therefore superior performances, was observed with thinner AEMs: beginning-of-life peak geometric power densities,  $P_{peak}$  of  $54 \rightarrow 94 \rightarrow 130 \text{ mW cm}^{-2}$  were recorded using  $154 \rightarrow 78 \rightarrow 51 \text{ }\mu\text{m}$  thick AEMs, respectively, under identical test conditions (less than  $20 \text{ mW cm}^{-2}$  was obtained with the  $230 \text{ }\mu\text{m}$  thick Tokuyama AHA AEM – not developed for use in fuel cells) (Fig. 2). The commercial Pt/C (20% mass –  $0.5 \text{ mg}_{Pt} \text{ cm}^{-2}$  loading)-containing Type A carbon cloth prefabricated cathodes (E-Tek division of BASF) that were used contained PTFE binder (current commercially available fuel cell electrodes are optimized for use with PEMs and not for AEMs) and it was suspected that this restricted the transport of the  $H_2O$  supplied in the humidified oxygen supply to the reaction sites.



**Alkaline Membrane Fuel Cells. Figure 2**

Variation in power density for  $H_2/O_2$  AMFCs as a function of fully hydrated RG-AEM thickness. The cells contained ETFE-derived RG-AEMs. All electrodes were treated with SION1 ionomer and utilized E-TekPt/C (20% mass) catalyst-loaded carbon cloth (squares) or carbon paper (diamonds) supports

Achieving an optimal MEA is still hindered by the available anionomers. Switching to the use of Toray carbon paper commercial electrodes ( $435 \text{ }\mu\text{m}$  thick containing Pt/C (20% mass) catalyst at  $0.5 \text{ mg}_{Pt} \text{ cm}^{-2}$  loading and containing PTFE binder – supplied by E-Tek division of BASF) at both the anode and cathode led to an increase in obtainable power densities (Fig. 2).  $P_{peak} = 230 \text{ mW cm}^{-2}$  and a maximum  $i$  at full discharge of  $1.3 \text{ A cm}^{-2}$  was obtained in a  $H_2/O_2$  fuel cell at  $50^\circ\text{C}$  with S20 ( $20 \text{ }\mu\text{m}$ ). A  $P_{peak} = 260 \text{ mW cm}^{-2}$  was obtained at  $50^\circ\text{C}$  with Tokuyama A201 using SION1-treated electrodes. Fully satisfactory MEA lamination cannot, however, be achieved with SION1 alkaline ionomer and there are still considerable, variable, contact resistances.

Pt-free electrodes are viable in AMFCs (more evidence is given below). Ag/C (60% mass,  $4 \text{ mg}_{Ag} \text{ cm}^{-2}$ ) performed well when used as a cathode catalyst (Ag/C catalyst performed poorly in the cathode of a PEMFC). Open circuit voltages ( $V_{OC}$ ) were normally in the range 1.0–1.1 V with Solvay ADP and Pt/C electrodes (ca. 100 mV higher than obtained with Nafion-115 PEMFCs); a drop of 100 mV in  $V_{OC}$  was observed when Pt was replaced with Ag at the cathode.

The activity of the carbon supports needs to be evaluated for each catalyst material to be tested. An important finding was that metal-free Vulcan XC-72R (i.e., carbon black only) cathodes produced significant currents ( $P_{peak} = 22 \text{ mW cm}^{-2}$  and  $V_{OC} = 0.88 \text{ V}$  with Solvay ADP at  $50^\circ\text{C}$ ). Carbons reduce  $O_2$  primarily through a  $2e^-$  mechanism at high pH producing hydroperoxide anions ( $HO_2^-$ ), which may form highly reactive hydroxyl radical species.

Other groups have been investigating nonradiation-grafted AEMs. A fully non-PGM-catalyst-containing metal-cation-free alkaline membrane fuel cell was first reported by the team at Wuhan University [73], for an AMFC containing a QA-polysulfone AEM and an MEA that did not contain any PGM catalysts; a  $P_{peak} = 52 \text{ mW cm}^{-2}$  at  $60^\circ\text{C}$  (open circuit voltage, OCV,  $V_{OC} = 0.94 \text{ V}$ ) was obtained using Ni-based anode catalyst and a Ag cathode catalyst.

Polysulfone-based AEMs (see earlier) have been tested in  $H_2/\text{air}$  fuel cells at  $60^\circ\text{C}$  using a polysulfone alkaline ionomer (dew point of the gases was  $55^\circ\text{C}$ ). A  $P_{peak} = 28 \text{ mW cm}^{-2}$  at  $V_{cell} = 0.47 \text{ V}$  was obtained



with Pt/C (40% mass, Johnson Matthey) with loadings of  $0.5 \text{ mg}_{\text{Pt}} \text{ cm}^{-2}$ .  $P_{\text{peak}} = 30 \text{ mW cm}^{-2}$  at  $V_{\text{cell}} = 0.42 \text{ V}$  was obtained when the cathode contained a Ag/C catalyst (40% mass, E-Tek) at a loading of  $2 \text{ mg}_{\text{Ag}} \text{ cm}^{-2}$ . OCVs were 100 mV lower with Ag compared to Pt, which correlates with the results above. A  $P_{\text{peak}} = 365 \text{ mW cm}^{-2}$  at 0.4 V ( $V_{\text{OC}} = 1.04 \text{ V}$ ,  $r = 0.152 \Omega \text{ cm}^2$ ) was reported at the 214th Meeting of the Electrochemical Society in 2008 by Wang's team at Penn State University, using Tokuyama's A901 ( $10 \mu\text{m}$ ) AEM and AS-4 alkaline ionomer ( $\text{IEC} = 1.3 \text{ meq g}^{-1}$ ) at  $50^\circ\text{C}$  in a  $\text{H}_2/\text{O}_2$  AMFC with  $0.4 \text{ mg cm}^{-2}$  Pt/C catalysts; these values dropped to  $212 \text{ mW cm}^{-2}$  ( $V_{\text{OC}} = 1.02 \text{ V}$ ,  $r = 0.19 \Omega \text{ cm}^2$ ) when the  $\text{O}_2$  at the cathode was replaced with purified air ( $<1 \text{ ppm CO}_2$ ) and the  $P_{\text{peak}}$  dropped to  $113 \text{ mW cm}^{-2}$  ( $r = 0.22 \Omega \text{ cm}^2$ ). The higher value represents the highest  $\text{H}_2/\text{O}_2$  AMFC performance publicly reported to date.

KOH-doped membranes (see earlier) represent metal-cation-containing AEMs. A  $\text{H}_2/\text{O}_2$  fuel cell has been tested with a KOH-doped PBI membrane ( $t = 40 \mu\text{m}$ ) at  $50^\circ\text{C}$  yielding an  $i = 620 \text{ mA cm}^{-2}$  at  $V_{\text{cell}} = 0.60 \text{ V}$  ( $P = 370 \text{ mW cm}^{-2}$ ) [60]. This performance was similar to that of the same apparatus with a Nafion-117 PEM and also, remarkably, with a PBI AEM doped with both KOH and  $\text{K}_2\text{CO}_3$ . The question of the long-term stability and leaching out of the KOH (if any) remains to be fully addressed for this class of AEM, as does the issue of potential for carbonate/bicarbonate precipitation.

Gu and Yan et al. [76] developed an ionomer, tris(2,4,6-trimethoxyphenyl) polysulfone-methylene quaternary phosphonium hydroxide (TPQPOH), which is soluble in low boiling point and water-soluble solvents such as methanol, ethanol, and *n*-propanol. The solubility of this kind of ionomer could be ideal for use in AMFC MEAs;  $196 \text{ mW cm}^{-2}$  was achieved at  $80^\circ\text{C}$  ( $\text{H}_2/\text{O}_2$  gases supplied at 250 kPa back-pressure) using this phosphonium ionomer with a  $70 \mu\text{m}$  thick FAA commercial membrane (Fumatech GmbH) as the AEM. A maximum power of more than  $250 \text{ mW cm}^{-2}$  was obtained in a  $\text{H}_2/\text{O}_2$  fuel cell at  $50^\circ\text{C}$  with a  $50 \mu\text{m}$  AEM and gas back-pressures of 250 kPa. The presence of methoxy ( $-\text{OCH}_3$ ) groups on the aromatic rings that are connected to the phosphonium ion centers seems essential to adequate chemical stability.

## The Application of AEMs in Other Fuel Cell Types

The utility of AEMs as potential electrolytes in fuel cells arises not only from the prospects for the use of non-Pt-group metal (non-PGM) catalysts and cheaper fuel cell components (less corrosive environment) but also from potential for use of alternative fuels.

Alcohols and diols, sodium borohydride ( $\text{NaBH}_4$ ), and hydrazine ( $\text{H}_2\text{NNH}_2$ ) have all been used directly as fuels in AMFCs. Patents [77, 78] introduced more fuels including hydrated hydrazine ( $\text{NH}_2\text{NH}_2 \cdot \text{H}_2\text{O}$ ), hydrazine carbonate ( $(\text{NH}_2\text{NH}_2)_2\text{CO}_2$ ), hydrazine sulfate ( $\text{NH}_2\text{NH}_2 \cdot \text{H}_2\text{SO}_4$ ), monomethyl hydrazine ( $\text{CH}_3\text{NHNH}_2$ ), ammonia ( $\text{NH}_3$ ), heterocycles such as imidazole and 1,3,5-triazine and 3-amino-1,2,4-triazole, and hydroxylamines such as hydroxylamine ( $\text{NH}_2\text{OH}$ ) and hydroxylamine sulfate ( $\text{NH}_2\text{OH} \cdot \text{H}_2\text{SO}_4$ ); the catalysts were co-based for the fuel side (anode) and Ag/C, Pt/C and Ni/C for the oxygen reduction side (cathode). EP2133946 [79] disclosed the use of transition metals as catalysts in AMFCs. There are numerous reports on the use of non-Pt catalyst, such as  $\text{MnO}_2$  [58], Ag/C [74, 75, 79], Au/C [80], FeTPP/BPC (Black Pearl Carbon) [81], CoPPyC [82], FeCo-CNF (Carbon Nanofiber) [83], CoFeN/C-HLH [84] in AMFC cathodes and Cr-decorated Ni/C [73], Fe-Co-Ni/C [85, 86] at anodes.

Zhuang et al. conducted a feasibility analysis of the use of AEMs in DMFCs [87]; this important study concentrated on the thermodynamic disadvantages versus kinetic advantages of such cells by considering the reactions taking place in aqueous solutions and the authors concluded that the thermodynamic voltage losses due to the pH difference across the AEM would be approximately canceled out by the kinetic voltage gains. Ogumi et al. studied direct alcohol fuel cells (methanol and ethylene glycol) with a commercial AEM (Tokoyama AHA, thickness  $\approx 240 \mu\text{m}$ ) [39]. Solutions of ethylene glycol and methanol were made to a concentration of  $1 \text{ mol dm}^{-3}$  dissolved in KOH (aq,  $1 \text{ mol dm}^{-3}$ ), and so the system was not metal-cation-free (common with other AMFCs obtaining reasonable power levels with alcohols as fuels). Cell voltages were around 100 mV higher for ethylene glycol compared to methanol. This fuel cell is *not* quite an alkaline analogue of an acid membrane DMFC, as the authors used alkaline fuel solution. No indication was

given concerning completeness of oxidation of the ethylene glycol to  $\text{CO}_2$  and  $\text{H}_2\text{O}$ . Yu and Scott also reported the operation of a direct methanol alkaline fuel cell, with platinized titanium mesh anodes ( $1.5 \text{ mg cm}^{-2}\text{Pt/Ti}$ ) and a commercial alkaline anion-exchange membrane (Morgane ADP membrane from Solvay SA) [88].

It has been proposed that ammonia would make a good energy vector/carrier [6] and an indirect fuel for a hydrogen fuel cell; ammonia is environmentally benign (it is already used as a fertilizer and neutralizes acid rain), already available cheaply, contains 50% more H per  $\text{dm}^3$  than liquid hydrogen, and is a liquid at much lower pressures (8–9 bar); it also has the advantage of a strong smell, allowing the easier location of leaks. Cracking ammonia to form hydrogen free of traces of ammonia for use in traditional PEMFCs requires temperatures of above  $900^\circ\text{C}$ ; PEMFCs cannot tolerate any  $\text{NH}_3$  contamination. An advantage of AFCs is that they will tolerate low levels of  $\text{NH}_3$ , and so cracking can be undertaken at lower temperatures; this has yet to be explored using AMFC approaches.

### A Brief Overview of the Patent Literature

Zeng and Varcoe have recently reviewed in depth the patent literature concerning AEMs [89]. A small number of patents of direct interest to application of AEMs in AMFCs are considered in this section.

In the area of novel AEMs:

- Patents relating to the application of radiation-grafted ion-exchange membranes in fuel cells have been granted to Scherer et al. [89] and to Stone and Steck [90, 91]. These patents mention the functionalization of base polymers with quaternary ammonium groups to yield alkaline polymers. The use of fluorine-substituted styrenic monomers is also claimed to improve membrane chemical stability when utilized in fuel cells (removal of undesired and reactive C–H bonds).
- Fauvarque was granted a patent on aqueous alkaline solid electrolyte based around polyether polymer matrices [92].
- Yao et al. were granted related patents on alkaline polymers containing alkyl quaternary ammonium salts, nitrogen containing heterocyclic quaternary

ammonium salts, and metal hydroxide salts for potential applications in alkaline batteries and fuel cells [93–95].

In the area of device architecture:

- Jaouen patented a novel cathode structure for a solid polymer fuel cell [96]; that cathode structure consisted of a solid polymer anion-exchange membrane surrounding the catalysts particles (entirely within the cathode structure), which was in contact (surrounded) with a cation-conducting polymer membrane.
- Lu and Zhuang were granted a patent on alkaline-resin-containing fuel cells with direct fuelling with liquid organic fuels [97].
- Related patents from Divisek describe the invention of a methanol fuel cell utilizing an anion-exchange membrane [98, 99].

### Future Directions and Research Challenges

There are a number of research challenges that must be overcome before AEMs can be successfully applied in pre-commercial AMFCs:

1. As a priority and to complement ex situ stability measurements, AEMs must be evaluated in  $\text{H}_2/\text{O}_2$ ,  $\text{H}_2/\text{air}$ , and methanol/air fuel cells for several thousand hours to ensure adequate in situ membrane stability and assess the effect of and level of  $\text{CO}_2^{2-}/\text{HCO}_3^-$  formation. The effect of adding alkali to the methanol solution must also be thoroughly investigated over long timescales.
2. AEMs must be created with higher conductivities, to ensure good fuel cell performances at high current densities. It should be noted, however, that with application of an alkaline DMFC as a power source for portable devices, an overriding priority is maximizing the energy density of the fuel and fuel efficiency (cell voltage). These DMFCs will be operated at low-to-medium current densities (higher efficiencies), where electrode overpotentials are the dominant cause of voltage losses.
3. It is essential that a solubilized form of an alkaline anion-exchange polymer be developed to improve the interface between the electrodes and the AEM electrolyte. Success in this effort will decrease MEA resistances. A water-based soluble form which can

be rendered water insoluble when cast would be preferred, as there are safety concerns (primarily with industrial scale production) about using organic solvents near finely dispersed (pyrophoric) metal catalysts (unsupported or supported on carbon).

4. If AEMs are to be applied to fuel cells for other applications (such as automotive power etc.), more temperature stable AEMs must be developed. Operation of alkaline-membrane-based fuel cells at elevated temperature would reduce thermodynamic voltage losses due to pH differences across the AEM and would also improve the electrokinetics. Successful, stable over the long term, and being operational at elevated temperatures, all would allow application in fuel cells for the automotive mass market.
5. If AEMs are shown to be stable in fuel cells over thousands of hours, an in-depth investigation into effective and cheaper non-noble metal catalysts (e.g., Ni, Ag etc.) is indicated. There would also be a greater chance of finding methanol-tolerant catalysts for use in the cathodes than in related PEM-based DMFCs.
6. Finally, removal of fluorine from the polymer systems would be of interest to enhance the environmental credentials of the technology (facilitating easier disposal) and reduce costs. The substitution of fully fluorinated FEP with non-fluorinated LDPE is feasible. While oxidative radical degradation is a problem at the cathode and anode (via oxygen diffusion through the membrane) with non-fluorinated PEMs, it has been shown using electron paramagnetic resonance (EPR) studies that such degradation is prevented in highly alkaline (pH > 11.7) conditions pertinent to AEMs [100–102].

## Acknowledgment

We thank the Engineering and Physical Sciences Research Council (EPSRC) of the United Kingdom for contracts with our team in the area of development of alkaline membrane technology for alkaline membrane fuel cells and associated electrical energy generation: GR/S60709/01, EP/F027524/1, EP/G009929/2 and EP/H025340/1.

## Bibliography

### Primary Literature

1. Varcoe JR, Slade RCT (2005) Prospects for alkaline anion-exchange membranes in low temperature fuel cells. *Fuel Cells* 5:189–200
2. Zeng R, Varcoe JR (2011) Alkaline anion exchange membranes for fuel cells – a patent review. *Recent Pat Chem Eng* 4:93–115
3. Schulze M, Gülzow E (2004) Degradation of nickel anodes in alkaline fuel cells. *J Power Sources* 127:252–263
4. Wagner N, Schulze M, Gülzow E (2004) Long term investigations of silver cathodes for alkaline fuel cells. *J Power Sources* 127:264–272
5. McLean GF, Niet T, Prince-Richard S, Djilali N (2002) An assessment of alkaline fuel cell technology. *Int J Hydrogen Energy* 27:507–526
6. Cifrain M, Kordesch KV (2004) Advances, aging mechanism and lifetime in AFCs with circulating electrolytes. *J Power Sources* 127:234–242
7. Gülzow E, Schulze M (2004) Long-term operation of AFC electrodes with CO<sub>2</sub> containing gases. *J Power Sources* 127:243–251
8. Zeng R, Poynton SD, Kizewski JP, Slade RCT, Varcoe JR (2010) A novel reference electrode for application in alkaline polymer electrolyte membrane fuel cells. *Electrochem Commun* 12:823–835
9. Yanagi H, Fukuta K (2008) Anion exchange membrane and ionomer for alkaline membrane fuel cells (AMFCs). *Electrochem Soc Trans* 16:257–262
10. Kizewski JP, Mudri NH, Zeng R, Poynton SD, Slade RCT, Varcoe JR (2010) Alkaline electrolytes and reference electrodes for alkaline polymer electrolyte membrane fuel cells. *Electrochem Soc Trans* 33:27–35
11. Yan JL, Hickner MA (2010) Anion exchange membranes by bromination of benzylmethyl-containing poly(sulfone)s. *Macromolecules* 43:2349–2356
12. Adams LA, Poynton SD, Tamain C, Slade RCT, Varcoe JR (2008) A carbon dioxide tolerant aqueous–electrolyte–free anion–exchange membrane alkaline fuel cell. *Chem Sus Chem* 1:79–81
13. Zagorodni AA, Kotova DL, Selemenev VF (2002) Infrared spectroscopy of ion exchange resins: chemical deterioration of the resins. *React Funct Polym* 53:157–171
14. Neagu V, Bunia I, Plesca I (2000) Ionic polymers – VI. Chemical stability of strong base anion exchangers in aggressive media. *Polym Degrad Stab* 70:463–468
15. Sata T, Tsujimoto M, Yamaguchi T, Matsusaki K (1996) Change of anion exchange membranes in an aqueous sodium hydroxide solution at high temperature. *J Membr Sci* 112:161–170
16. Soda T, Kaisha K (1972) Anion exchange membranes and their production. GB Patent 1401997
17. Hansen RD, Wheaton RM (1966) Separation of acid from polymers by dialysis with anion exchange membrane. US Patent 3244620

18. Imoto R, Kosaka Y, Shimizu A (1966) Process for manufacturing anion-exchange membranes from a graft copolymer of SBR and a vinylpyridine reacted with an epoxy resin. US Patent 3258435
19. Hansen RD, Wheaton RM (1966) Separation of acid by dialysis with anion-exchange membranes. US Patent 3272737
20. Süser A (1973) Preparation of anion exchange membranes from cellulose sheets. US Patent 3714010
21. Scott, Rosedale, Moulton, Taylor, Gough (1990) Anion exchange membranes. Patent EP0382439
22. Altmeier P (1995) Stark basische anionenaustauscher-membranen und verfahren zu deren herstellung. Patent WO9506083
23. Altmeier P (1998) Strongly alkaline anion exchange membranes and process for producing the same. US Patent 5746917
24. Ehrikovich KJ, Viktorovna SN, Aleksandrovich FJ, Semenovitch GS, Mikhajlovich AJ, Vladimirovna SI, Alekseevna JN, Fedorovich TS (1997) Method for production of anion-exchange membranes having high penetrability to chloride ion. Patent RU2074204
25. Jurevich TD (2007) Method for modifying anion-exchange membranes MA-40. Patent RU2303835
26. Aminabhavi T, Kulkarni PV, Kariduraganavar MY (2009) Ion exchange membranes, methods and processes for production thereof and uses in specific applications. US Patent 7544278
27. Danks TN, Slade RCT, Varcoe JR (2002) Comparison of PVDF- and FEP-based radiation-grafted alkaline anion-exchange membranes for use in low temperature portable DMFCs. *J Mater Chem* 12:3371–3373
28. Danks TN, Slade RCT, Varcoe JR (2003) Alkaline anion-exchange radiation-grafted membranes for possible electrochemical application in fuel cells. *J Mater Chem* 13:712–721
29. Herman H, Slade RCT, Varcoe JR (2003) The radiation-grafting of vinylbenzyl chloride onto poly(hexafluoropropylene-co-tetrafluoroethylene) films with subsequent conversion to alkaline anion-exchange membranes: optimisation of the experimental conditions and characterization. *J Membr Sci* 218:147–163
30. Tzanetakos N, Varcoe J, Slade RS, Scott K (2003) Salt splitting with radiation grafted PVDF anion-exchange membrane. *Electrochem Commun* 5:115–119
31. Tzanetakos N, Varcoe JR, Slade RCT, Scott K (2005) Radiation-grafted PVDF anion exchange membrane for salt splitting. *Desalination* 174:257–265
32. Slade RCT, Varcoe JR (2005) Investigations of conductivity in FEP-based radiation-grafted alkaline anion-exchange membranes. *Solid State Ionics* 176:585–597
33. Varcoe JR, Slade RCT (2006) An electron-beam-grafted ETFE alkaline anion-exchange membrane in metal-cation-free solid-state alkaline fuel cells. *Electrochem Commun* 8:839–843
34. Varcoe JR, Slade RCT, Yee ELH, Poynton SD, Driscoll DJ, Apperley DC (2007) Poly(ethylene-co-tetrafluoroethylene)-derived radiation-grafted anion-exchange membrane with properties specifically tailored for application in metal-cation-free alkaline polymer electrolyte fuel cells. *Chem Mater* 19:2686–2693
35. Chempath S, Einsla BR, Pratt LR, Macomber CS, Boncella JM, Rau JA, Pivovar BS (2008) Mechanism of tetra alkyl ammonium head group degradation in alkaline fuel cell membranes. *J Phys Chem C* 112:3179–3182
36. Piana M, Boccia M, Filpi A, Flammia E, Miller HA, Orsini M, Salusti F, Santiccioloi S, Ciardelli F, Pucci A (2010) H<sub>2</sub>/air alkaline membrane fuel cell performance and durability, using novel ionomer and non-platinum group metal cathode catalyst. *J Power Sources* 195:5875–5881
37. Li YS, Zhao TS, Liang ZX (2009) Performance of alkaline electrolyte-membrane-based direct ethanol fuel cells. *J Power Sources* 187:387–392
38. Kim JH, Kim HK, Hwang KT, Lee JY (2010) Performance of air-breathing direct methanol fuel cell with anion-exchange membrane. *Int J Hydrogen Energy* 35:768–773
39. Matsuoka K, Iriyama Y, Abe T, Matsuoka M, Ogumi Z (2005) Alkaline direct alcohol fuel cells using an anion exchange membrane. *J Power Sources* 150:27–31
40. Li YS, Zhao TS, Liang ZX (2009) Effect of polymer binders in anode catalyst layer on performance of alkaline direct ethanol fuel cells. *J Power Sources* 190:223–229
41. Fujiwara N, Siroma Z, Yamazaki SI, Ioroi T, Senoh H, Yasuda K (2008) Direct ethanol fuel cells using an anion exchange membrane. *J Power Sources* 185:621–626
42. Bunazawa H, Yamazaki Y (2008) Influence of anion ionomer content and silver cathode catalyst on the performance of alkaline membrane electrode assemblies (MEAs) for direct methanol fuel cells (DMFCs). *J Power Sources* 182:48–51
43. Marx D, Chandra A, Tuckerman ME (2010) Aqueous basic solutions: hydroxide solvation, structural diffusion, and comparison to the hydrated proton. *Chem Rev* 110:2174–2216
44. Hibbs MR, Hickner MA, Alam TM, McIntyre SK, Fujimoto CH, Cornelius CJ (2008) Transport properties of hydroxide and proton conducting membranes. *Chem Mater* 20:2566–2573
45. Robertson NJ, Kostalik HA IV, Clark TJ, Mutolo PF, Abruna HD, Coates GW (2010) Tunable high performance cross-linked alkaline anion exchange membranes for fuel cell applications. *J Am Chem Soc* 132:3400–3404
46. Tanaka M, Koike M, Miyatake K, Watanabe M (2010) Anion conductive aromatic ionomers containing fluorenyl groups. *Macromolecules* 43:2657–2659
47. Wang JH, Li SH, Zhang SB (2010) Novel hydroxide-conducting polyelectrolyte composed of a poly(arylene ether sulfone) containing pendant quaternary guanidinium groups for alkaline fuel cell applications. *Macromolecules* 43:3890–3896
48. Guo ML, Fang J, Xu HK, Li W, Lu XH, Lan CH, Li KY (2010) Synthesis and characterization of novel anion exchange membranes based on imidazolium-type ionic liquid for alkaline fuel cells. *J Membr Sci* 362:97–104
49. Wang JH, Zhao Z, Gong FX, Li SH, Zhang SB (2009) Synthesis of soluble poly(arylene ether sulfone) ionomers with pendant quaternary ammonium groups for anion exchange membranes. *Macromolecules* 42:8711–8717

50. Tripathi BP, Kumar M, Shahi VK (2010) Organic–inorganic hybrid alkaline membranes by epoxide ring opening for direct methanol fuel cell applications. *J Membr Sci* 360:90–101
51. Clark TJ, Robertson NJ, Kostalik HA IV, Lobkovsky EB, Mutolo PF, Abruna HD, Coates GW (2009) A ring–opening metathesis polymerization route to alkaline anion exchange membranes: development of hydroxide–conducting thin films from an ammonium–functionalized monomer. *J Am Chem Soc* 131: 12888–12889
52. Shevchenko VV, Gumennaya MA (2010) Synthesis and properties of anion-exchange membranes for fuel cells. *Theor Exp Chem* 46:139–152
53. Varcoe JR, Poynton SD, Slade RCT (2009) In: Vielstich W, Gasteiger HA, Yokokawa H (eds) *Handbook of fuel cells – fundamentals, technology and applications*, vol 5, *Advances in electrocatalysis, materials, diagnostics and durability*. Wiley, Chichester, pp 322–336
54. Zhou J, Unlu M, Vega J, Kohl P (2009) Anionic polysulfone ionomers and membranes containing fluorenyl groups for anionic fuel cells. *J Power Sources* 190:285–292
55. Fang J, Shen PK (2006) Quaternized poly(phthalazinone ether sulfone ketone) membrane for anion exchange membrane fuel cells. *J Membr Sci* 285:317–322
56. Zeng QH, Liu QL, Broadwell I, Zhu AM, Xiong Y, Tu XP (2010) Anion exchange membranes based on quaternized polystyrene–block–poly(ethylene–ran–butylene)–block–polystyrene for direct methanol alkaline fuel cells. *J Membr Sci* 349:237–243
57. Hou HY, Sun GQ, He RH, Wu ZM, Sun BY (2008) Alkali doped polybenzimidazole membrane for high performance alkaline direct ethanol fuel cell. *J Power Sources* 182:95–99
58. Hou HY, Sun GQ, He RH, Sun BY, Jin W, Liu H, Xin Q (2008) Alkali doped polybenzimidazole membrane for alkaline direct methanol fuel cell. *Int J Hydrogen Energy* 33:7172–7176
59. Modestov AD, Tarasevich MR, Leykin AY, Filimonov VY (2009) MEA for alkaline direct ethanol fuel cell with alkali doped PBI membrane and non–platinum electrodes. *J Power Sources* 188:502–506
60. Xing B, Savadogo O (2000) Hydrogen oxygen polymer electrolyte membrane fuel cells (PEMFCs) based on alkaline–doped polybenzimidazole (PBI). *Electrochem Commun* 2:697–702
61. Savadogo O (2004) Emerging membranes for electrochemical systems. Part II. High temperature composite membranes for polymer electrolyte fuel cell (PEFC) applications. *J Power Sources* 127:135–161
62. Yang CC, Lin CT, Chiu SJ (2008) Preparation of the PVA/HAP composite polymer membrane for alkaline DMFC application. *Desalination* 233:137–146
63. Yang CC, Chiu SJ, Chien WC, Chiu SS (2010) Quaternized poly(vinyl alcohol)/alumina composite polymer membranes for alkaline direct methanol fuel cells. *J Power Sources* 195:2212–2219
64. Yang CC, Chiu SJ, Lee KT, Chien WC, Lin CT, Huang CA (2008) Study of poly(vinyl alcohol)/titanium oxide composite polymer membranes and their application on alkaline direct alcohol fuel cell. *J Power Sources* 184:44–51
65. Yang CC (2007) Synthesis and characterization of the cross–linked PVA/TiO<sub>2</sub> composite polymer membrane for alkaline DMFC. *J Membr Sci* 288:51–60
66. Wan Y, Peppley B, Creber KAM, Bui VT, Halliop E (2006) Preliminary evaluation of an alkaline chitosan–based membrane fuel cell. *J Power Sources* 162:105–113
67. Wan Y, Creber KAM, Peppley B, Bui VT (2006) Chitosan–based electrolyte composite membranes II. Mechanical properties and ionic conductivity. *J Membr Sci* 284:331–338
68. Wan Y, Creber KAM, Peppley B, Bui VT (2006) Chitosan–based solid electrolyte composite membranes I. Preparation and characterization. *J Membr Sci* 280:666–674
69. Vassal N, Salmon E, Fauvarque JF (2000) Electrochemical properties of an alkaline solid polymer electrolyte based on P(ECH–co–EO). *Electrochim Acta* 45:1527–1532
70. Yang CC, Chiu SJ, Chien WC (2006) Development of alkaline direct methanol fuel cells based on cross–linked PVA polymer membranes. *J Power Sources* 162:21–29
71. Fauvarque JF (1996) Alkaline solid polymer electrolyte, electrode and electrochemical generator containing such as electrolyte. US Patent 5569559
72. Varcoe JR, Slade RCT, Lam How Yee E (2006) An alkaline polymer electrochemical interface: a breakthrough in application of alkaline anion–exchange membranes in fuel cells. *Chem Commun* :1428–29
73. Lu SF, Pan J, Huang AB, Zhuang L, Lu JT (2008) Alkaline polymer electrolyte fuel cells completely free from noble metal catalysts. *Proc Natl Acad Sci* 105:20611–20614
74. Park JS, Park SH, Yim SD, Yoon YG, Lee WY, Kim CS (2008) Performance of solid alkaline fuel cells employing anion–exchange membranes. *J Power Sources* 178:620–626
75. Tang DP, Pan J, Lu SF, Zhuang L, Lu JT (2010) Alkaline polymer electrolyte fuel cells: principle, challenges, and recent progress. *Sci China Chem* 53:357–364
76. Gu S, Cai R, Luo T, Chen ZW, Sun MW, Liu Y, He GH, Yan YS (2009) A soluble and highly conductive ionomer for high–performance hydroxide exchange membrane fuel cells. *Angew Chem Int Ed* 48:6499–6502
77. Asazawa K, Yamada K, Tanaka H (2007) Fuel cell. Patent EP1843416
78. Tanaka H, Yamada K, Asazawa K (2004) Fuel cell. Patent EP1460705
79. Asazawa K, Yamada K, Tanaka H (2009) Fuel cell. Patent EP2133946
80. Poynton SD, Kizewski JP, Slade RCT, Varcoe JR (2010) Novel electrolyte membranes and non–Pt catalysts for low temperature fuel cells. *Solid State Ionics* 181:219–222
81. Ramaswamy N, Mukerjee S (2010) Electrocatalysis of oxygen reduction on non–precious metallic centers at high pH environments. *ECS Trans* 33:1777–1785
82. Asazawa K, Yamamoto K, Yamada K, Tanaka H, Matsumura D, Tamura K, Nishihata Y, Atanassov P (2010) XAFS analysis of unpyrolyzed CoPPyC oxygen reduction catalyst for anion–exchange membrane fuel cells (AMFC). *Electrochem Soc Trans* 33:1751–1755



83. Jeong B, Uhm S, Lee J (2010) Iron-cobalt modified electrospun carbon nanofibers as oxygen reduction catalysts in alkaline fuel cells. *Electrochem Soc Trans* 33:1757–1767
84. Li XG, Popov BN, Kawahara T, Yanagi H (2010) Recent advances in non-precious metal catalysts for oxygen reduction reaction in fuel cells. *Electrochem Soc Trans* 33:1769–1776
85. Takeguchi T, Takahashi H, Yamanaka T, Nakamura A, Ueda W (2010) Development of direct-ethanol anion-conducting solid alkaline inorganic fuel cell. *Electrochem Soc Trans* 33:1847–1851
86. Nakamura A, Takahashi H, Takeguchi T, Yamanaka T, Wang Q, Uchimoto Y, Ueda W (2010) Effect of reduction temperature of Fe-Co-Ni/C catalyst on the solid alkaline fuel cell performance. *Electrochem Soc Trans* 33:1817–1821
87. Wang Y, Li L, Hu L, Zhuang L, Lu J, Xu B (2003) A feasibility analysis for alkaline membrane direct methanol fuel cell: thermodynamic disadvantages versus kinetic advantages. *Electrochem Commun* 5:662–666
88. Yu EH, Krewer U, Scott K (2010) Principles and materials aspects of direct alkaline alcohol fuel cells. *Energies* 3:1499–1528
89. Scherer GG, Büchi FN, Gupta G (Paul Scherrer Institut, Switzerland) (1997) US Patent 5656386
90. Stone C, Steck A (Ballard Power Systems Inc., Canada) (2002) US Patent 6359019
91. Stone C, Steck A (Ballard Power Systems Inc., Canada) (2000) PCT Patent 01/58576
92. Fauvarque J-F (CNAM Paris and Electricité de France) (1996) US Patent 5569559
93. Yao W, Tsai T, Chang Y-M, Chen M (Reveo Inc., USA) (2001) US Patent 6183914
94. Yao W, Tsai T, Chang Y-M, Chen M (Reveo Inc., USA) (2000) PCT Patent 00/16422
95. Yao W, Tsai T, Chang Y-M, Chen M (Reveo Inc., USA) (2000) Patent EP1116291
96. Jaouen F (ABB AB and Volvo, Sweden) (2002) PCT Patent 02/35633
97. Lu J, Zhuang L (Wuhan University) (2003) Patent CN1402370
98. Divisek J (Forschungszentrum Jülich, Germany) (2001) PCT Patent 01/61776
99. Divisek J (Forschungszentrum Jülich, Germany) (2001) Patent EP1256142
100. Yu J, Yi B, Xing D, Liu F, Shao Z, Fu Y, Zhang H (2003) Degradation mechanism of polystyrene sulfonic acid membrane and application of its composite membranes in fuel cells. *Phys Chem Chem Phys* 5:611–615
101. Hübner G, Roduner E (1999) EPR investigation of HQ<sup>•</sup> radical initiated degradation reactions of sulfonated aromatics as model compounds for fuel cell proton conducting membranes. *J Mater Chem* 9:409–418
102. Panchenko A, Dilger H, Möller E, Sixt T, Roduner E (2004) In situ EPR investigation of polymer electrolyte membrane degradation in fuel cell applications. *J Power Sources* 127:325–330

## Animal Breeding and Genetics, Introduction

IGNACY MISZTAL

Department of Animal and Dairy Science Breeding and Genetics, University of Georgia, Athens, GA, USA

Informal animal breeding started thousands of years ago when hunter-gatherers started domesticating animals. Out of thousands of species available only few were domesticated as the requirements for domestication were numerous: plant diet, fast growth rate, ability to breed in captivity, good disposition, little tendency to panic, and ability to function well in groups [1]. An ancient farmer/herder took special care of well-behaving animals that provided good growth, or plenty of milk, or reliable draft, or lots of wool, etc., while eliminating the troublesome ones. The domestication greatly increased the nutritional output per unit of land although it also brought new problems, e.g., new diseases and wars. These problems were smaller than the benefits of the domestication as hunter-gatherers mostly disappeared. Some of the benefits could be due to a positive effect of animal products on IQ [2].

The natural selection maximizes survival under the natural conditions. Under domestication, the selection maximizes utility of a specie for a farmer while deemphasizing and thus reducing energy expenditure for characteristics less important or unimportant under domestication, e.g., fighting ability to select a mate or defend against predators, ability to cover long distances, etc. [3]. The degree of economically beneficial selection is environment dependent because some traits (e.g., resistance to harsh conditions) may be necessary in some environments while they are redundant in other environments. In the end, the improved animals have a smaller environmental imprint [4].

In ► [Animal Breeding Methods and Sustainability](#), Agustín Blasco provides a historical perspective to animal breeding including the creation of breeds. Although the breeding has been practiced for many millennia, the science behind it is relatively recent. With new breeding tools, the genetic improvement accelerated resulting in much higher productivity per

animal with lower cost per unit of animal product. High productivity of a few improved breeds raises a question of survival and consequently conservation of less improved breeds. Also, highly improved breeds may not be optimal in more demanding environments.

In “► [Animal Breeding, Foundations of](#),” Guilherme Rosa focuses on theories that made modern animal breeding possible. These are population genetics, quantitative genetics, mixed models, and related issues. He examines the infinitesimal genetic model where it is assumed that a trait is controlled by a large number of independent loci, and he mentions models involving quantitative trait loci (QTL) or major genes. Selection can be for a single trait with possibly undesirable response for some of the remaining traits, or it can be multitrait, where weights on traits are economically derived for a more balanced breeding.

The selection in animal breeding depends on models and often sophisticated computing to estimate parameters of those models. In “► [Animal Breeding, Modeling in](#),” Lawrence Schaeffer presents mixed models that are commonly used to analyze many traits in small and large populations. In particular, an animal model considers all phenotypes and pedigrees to provide best unbiased linear predictions of animals’ breeding values (EBV). While the basic animal model may be sufficient for fairly accurate predictions, sometimes additional features are necessary to better reflect the complexity of data. Large differences in variability within the environments require a model that accounts for heterogeneous variances per environment, categorical traits are best analyzed by a threshold model; special models are needed to analyze data censored, e.g., by time.

While most of the progress in animal breeding was based on the infinitesimal model, the availability of genetic markers raised hopes of finding major genes. Subsequently, marker-assisted selection (MAS) would allow determining EBV for young animals without waiting for phenotypes. In his entry “► [Animal Molecular Genetics from major genes to genomics](#),” Asko Mäki-Tanila describes theories for finding markers or QTL and applications of MAS. While several large QTLs have been found, in general, the estimated values for large QTL seem to be inflated while many QTLs are below the detection level. Thus, the total contribution of large QTLs seems to be small for most traits. Currently, individuals can be genotyped inexpensively for

a large number of SNP markers. The use of these markers for breeding is called the genomic selection. The genomic selection is successful in populations with small effective population sizes where the genomic information provides more accurate relationships.

A selection is usually performed in a specific environment, which is defined as a combination of climate, nutrition, management, and stability. Animals excelling in one environment may perform poorly in another environment and vice versa, or show a genotype by environment interaction (GxE). In “► [Animal Genetic in Environment Interaction](#),” Erling Strandberg describes terms useful for quantification of GxE such as plasticity and homeostasis. GxE can be analyzed by reaction norms. We are interested in GxE only when genotypes rank differently in different environments.

The traditional selection is by selecting the “best” individuals. However, these individuals may indirectly be selected for poor group performance, e.g., aggression, and such animals may inhibit the growth of the other animals. In “► [Socially affected traits, Inheritance and genetic improvement](#),” Peter Bijma presents cases where the group performance is more important than the individual performance, proposes models that can be used to breed “socially adapted” animals, and describes results of experiments with such breeding. The social performance is most important for animals bred in cages such as poultry and pigs where the aggression can result in physical harm or even death. An intuitively obvious solution to aggression of allowing more space per animal in fact increases fighting and subsequently causes economic losses, as extra space allows for more fighting.

The next few entries are devoted to breeding for specific species.

In the entry ► [Poultry Breeding](#), Yoav Eitan and Morris Soller document how, over the past 100 years, the chicken meat changed from being one of the most expensive to the least expensive. Such a progress required a good choice of initial breeds, intensive selection, and diligent research to discover sources of new problems and their mitigation, and constantly adapting management to address these problems. The success with chicken is possible only when the environment can be tightly controlled as highly adapted animals retain minimal flexibility to handle less than the optimal conditions. An important issue in chicken is

animal welfare. While some stresses are unavoidable, like in nature, reducing the avoidable distress can be good economics.

In the entry ► [Pig breeding for increased sustainability](#), Pieter Knapp examines broader issues in breeding using pigs as an example. The first issue is diversity. Should different breeds be preserved or is it unimportant? What is the optimal size of the breeding population to sustain genetic variability and maximize the genetic progress over the long run? Pigs have been singled out as causing environment pollutions. Can breeding minimize the pollution? One way to minimize pollution and environmental damage is to breed efficient animals with good feed conversion ratio (FCR). Can one radically improve FCR? Is animal welfare a liability or an asset? The entry by Pieter Knapp contains extensive references.

Filippo Miglior, Sarah Locker, and Roger Shanks take a look into dairy breeding in their entry ► [Dairy Cattle Breeding](#). An intensive selection for milk made Holsteins the highest producing and the most popular dairy breed in temperate countries. Most of the progress in Holsteins, and to a lesser intent, in other breeds, is through the extensive use of highly select sires through artificial insemination. The selection of sires is in fact global because national sire evaluation from some 30 countries are now pooled together by an international agency “Interbull,” and semen of bulls ranked by this agency are available worldwide. Strong selection for production had an undesirable effect on fitness, including reduced fertility and survival. This side effect is remedied now by increased emphasis on fertility and survival in the selection index, and by crossbreeding. Lately, genomic selection greatly changed the breeding scheme in dairy.

Matthew Spangler describes beef breeding in the USA in his entry ► [Breeding in Beef Cattle](#). As opposed to dairy, the beef population consists of many breeds, and most animals sold for beef are crossbreds. Beef is initially raised in ranches and later brought to feedlots for a short time for accelerated growth. Animals are genetically evaluated based on growth at a few age points, for carcass characteristics and for fertility. Important issues in beef evaluation are heterosis and recombination loss in crosses. FCR in beef is much lower than in poultry or pigs; however, beef can utilize land unsuitable for crops.

Thorvador Árnason, in his entry ► [Breeding in Horses](#), presents issues in breeding horses, predominantly for racing. Such breeding is successful as the trend for maximum speed is positive. As only selected individuals are run in races, breeding for speed requires accounting for censoring, i.e., lack of records for an important part of the population. This is done by treating as phenotype a categorical variable of “has” or “does not have” records. An issue specific for horses is increased inbreeding due to very intensive selection and small populations. Large increase in inbreeding has negative effects on many traits and also increases the chances of propagating a recessive gene.

SWP Cloete looks at aspects of ► [breeding in developing countries and tropics](#). Increased production due to the animal breeding and limited population growth was successful in developing countries in creating surpluses of animal products. In developing countries, the animal breeding was less successful while the population exploded, resulting in shortages of animal products. The breeding was less successful because of specific challenges in developing countries and especially the tropics, and much smaller R&D. Traits especially important in the tropics are disease resistance (including tick and trypanosome), draft resistance, and ability to produce under periodic feed shortages. Subsequently, animals bred in the developed world may not survive in the tropics; however, their crosses sometimes do well. This entry contains a large number of references.

Even though the animal breeding is successful, there is a question whether the current pace of progress can be maintained. For example, FCR cannot be decreased below some 1.5 kg feed/1 kg of meat, unless the extra selection results in increased water content. Also, increased milk production at a cost of reduced fitness (less fertility, lower survival, more susceptibility to diseases) can at one point decrease overall profitability. Peer Berg ponders long-term challenges in animal breeding in his entry titled ► [Animal Breeding, Long-term Challenges](#). Low effective population sizes make whole populations less biosecure. Too optimized genotypes may require huge facilities that could destroy the environment and rural life. Also, new requirements for welfare may require changing breeding goals. However, most long-term challenges are not well known.

## Bibliography

1. Diamond J (2005) Guns, germs and steel. W.W. Norton, New York, 518 pp
2. Neumann CG, Murphy SP, Gewa C, Grillenberger M, Bwibo NO (2007) Meat supplementation improves growth, cognitive, and behavioral outcomes in Kenyan children. *J Nutr* 137(4):1119–1123
3. De Jong G, Van Noordwijk AJ (1992) Acquisition and allocation of resources: genetic (co)variances, selection, and life histories. *Am Nat* 139:749–770
4. Von Braun J (2010) The role of livestock production for a growing world population. [http://www.lohmann-information.com/content/l\\_i\\_45\\_2010\\_10.pdf](http://www.lohmann-information.com/content/l_i_45_2010_10.pdf)

## Animal Breeding Methods and Sustainability

AGUSTIN BLASCO

Institute of Animal Science and Technology,  
Universidad Politécnica de Valencia, Valencia, Spain

## Article Outline

Glossary  
Definition of the Subject  
Introduction  
Animal Breeding and Sustainability  
Animal Breeding Methods and Schemes  
Future Directions  
Bibliography

## Glossary

**Breed** A population of animals with common morphological characteristics that is recognized as a breed by the administration, by a breeders association, or by other groups of people.

**Best linear unbiased prediction (BLUP)** It is the most common statistical method used in breeding evaluation.

**Quantitative trait locus (QTL)** A gene having influence in a quantitative trait.

**Markers** Fragments of the DNA molecule for which their position is known.

**Response to selection** Genetic progress.

## Definition of the Subject

After domestication, animals were selected in different environments and for different traits leading to the modern breeds. Long before the appearance of the science called now as “Genetics,” animal breeding had been practiced by humans following intuitive criteria, less efficient than the scientific ones, but criteria that had provided success along many generations of selection [1]. The lack of a theory explaining inheritance slowed down animal breeding for many years, but with the rediscovery of Mendel’s rules at the beginning of the twentieth century and the development of quantitative genetics in the 1920s and 1930s animal breeding had the tools needed for its development. Animal breeding methods were developed in the 1930s and 1940s, and the first animal breeding companies and cooperatives started in using scientific methods for animal selection [2]. The development of artificial insemination in cattle in the 1940s and frozen semen in the 1950s led to the modern schemes of progeny test, in which bulls are proved with a high number of daughters, and semen of the best bulls is available worldwide. Large companies of animal breeding were created in the 1960s for poultry and pigs, and nowadays they dominate the market of reproducers, particularly in the avian case. In 1953 the structure of the DNA was published, leading to a quick development of all molecular genetics techniques. Today, DNA information is widely used as a complementary tool to the statistical methods based on data from records, to estimate the genetic values of the candidates to selection [3]. Although the commerce of genes is now extended worldwide, there is a recent interest in conserving breeds in danger of extinction due to this globalization. These breeds are a genes reserve for ensuring possible changes in the future market. Besides, some breeds can be helpful for developing sustainable systems in areas in which modern developed animals cannot be bred because of the lack of resources, climate, or other reasons [4].

## Introduction

Long before the appearance of the science that is now called “Genetics,” animal breeding had been practiced by humans following intuitive criteria, less efficient than the scientific ones, but criteria that had provided

success along many generations of selection. Darwin himself was impressed by the achievements of farmers, and artificial selection was a source of inspiration for his theory of evolution [5].

- We cannot suppose that all the breeds were suddenly produced as perfect and as useful as we see now them; indeed, in several cases, we know that this has not been their history. The key is man's power of accumulative selection: nature gives successive variations; man adds them up in certain directions useful to him. In this sense he may be said to make for himself useful breeds.

C. DARWIN

On the origin of species (1859, p. 30)

Animal breeding starts with domestication. Although there are several theories about the domestication process, it is generally admitted that selective breeding led to modern domestic animals, a hypothesis corroborated by the experiment of domestication of wild silver foxes started by Dimitri Belyaev in 1959 and still continued. After 40 years of selection for quiet temperament, silver foxes, which are aggressive to humans in the wild, became as friendly as dogs [6]. As a correlated response, some physical appearance also changed, and some bones of the skull were modified in the same direction as dogs when compared with wolves [7]. Modern molecular techniques permit to reconstruct the history of domestication [8]. After domestication, animals were selected in different environments and for different traits, leading to the modern breeds. References to animal breeding can be found in ancient Greek and Roman authors' works [9]; however, modern breeding practices start with the self-taught work of Robert Bakewell (1725–1795), who produced new breeds and had a high reputation as breeder [1]. He focused his work in the performances of his cattle and sheep, hiring rams, recording the offspring and keeping the sons of the best males. He fixed few and clear breeding objectives mating the best females with the best males. However, he disregarded the damaging effects of inbreeding and due to this, he had fertility troubles with his new breeds, but he is still considered as the first farmer practicing modern animal breeding.

The lack of a theory explaining inheritance slowed down animal breeding for many years. The theory of blending inheritance, sustaining that offspring was intermediate between parents, could not explain the

persistence of genetic variability. Some hybrid breeders had noticed that crossing hybrids they can recover discrete traits that were present in the parental population [10], but Gregor Mendel (1822–1884) was the first in calculating the frequencies in which the observed traits were transmitted, allowing him to propose the first rules of genetic inheritance [11]. Although Mendel was conscious of the importance of his research, his work, published in a context of hybrid plant production, was largely ignored until it was rediscovered at the beginning of the twentieth century, and it was widely used to explain the inheritance of discrete observable traits. Mendel's rules worked well for discrete traits like yellow or green color, but many traits like milk production or body weight showed a continuous variation and seemed to follow different inheritance rules. The biometrician school, founded by Karl Pearson (1857–1936), was using and developing statistical methods, and rejected Mendel's rules, considering them as a special case of inheritance for some discrete characters [12]. Mendel was aware about the fact that the simple rules he discovered could not be applied to continuous variation, but he suggested that in these cases many inheritance factors might act simultaneously producing all intermediate indistinguishable classes. After some exam of this possibility, it was disregarded by the biometricians, and a bitter dispute about the mechanisms of inheritance started until Fisher (1890–1962), in a seminal paper [13], used statistical methods to reconcile Mendel's laws on inheritance with the continuous variation observed by biometricians. (For a history of early development of genetics and this dispute, see [12].)

The work of Fisher in this and subsequent papers started both modern statistics and modern quantitative genetics, but the methods of this new science had still to be applied to animal breeding. This task was accomplished by Lush (1896–1982), who harmonized breeding practices with the knowledge provided by the new discipline. Lush defined concepts like heritability, and proposed methods of selection including the information of relatives, weighed according to the genetic contribution predicted by Mendel's rules and quantitative genetics. The several editions of his book "Animal breeding plans" contributed to spread the new knowledge among scientists, technicians, and breeders [14]. Modern indexes of selection for several traits were



developed for plants by Fairfield Smith [15] closely following some indications given by Fisher, and Hazel [16] applied them to animal breeding allowing on one side to use family information and on the other side to weigh all traits of economic interest according to the predicted benefits that the offspring would give.

The development of artificial insemination permitted having offspring of the same sires in many farms (see [17] for a history of its development). As environmental effects were different depending on the farm circumstances, data had to be corrected in order to evaluate the animals properly. Corrections for environmental effects like parity, season, length of lactation, etc., had been made before, but then the problem was more complex. Several methods were developed to pre-correct the data before genetic analysis was made, but it was Henderson (1911–1989), who proposed a method for integrating the genetic values and the environmental ones in the same statistical model. This allowed the prediction of genetic values at the same time that corrections for environmental values were made [18]. The development of computers allowed using all relatives in the evaluation, and some computing difficulties derived from the use of all relatives were solved by Henderson himself [19]. Nowadays his method called best linear unbiased prediction (BLUP) is the standard method in animal breeding evaluation. BLUP needs the variance components (genetic and environmental) for predicting the genetic values. To estimate them is a difficult task, because data come from different farms and different environments and they should be corrected as before. Paterson and Thompson [20] showed how to correct for the environmental effects and how to estimate the genetic variance components at the same time. Their method is called REML (Residual or Restricted Maximum Likelihood) and it is a standard for variance component estimation.

Animal breeding was dominated by REML and BLUP – and they are still the most common methods – until the development of modern computers allowed the use of Bayesian methods. These methods use probabilities for inferences, which give them several advantages and permit to express the uncertainty about the unknowns in a natural way. For example, it is easier to understand that the probability of a breed having a higher growth rate than another is 93% than to understand that when estimating the difference in

growth rate between breeds, in an infinite number of repetitions of the experiment, new samples will be higher than the actual sample in a 7% of the cases (which is the definition of a *P*-value of 7%). A review of Bayesian methods compared with classical statistical methods in animal breeding can be found in [21]. Bayesian methods were introduced by Daniel Gianola in the 1980s [22], but they lead to complicated integrals that could not be solved even by approximate methods. The rediscovery of a numerical method called Monte Carlo Markov Chains permitted to overcome this problem and to use Bayesian methods, leading to a high development and extension of them in animal breeding (see [23] for a detailed exposition of the methods).

With the arrival of DNA analysis techniques, a new field was open for research. Transgenesis looked as a promising area, but its real usefulness in animal breeding has been discussed [24]. Molecular markers, however, have been widely used in animal breeding as a complementary tool in genetic programs. They have been also used for capturing major genes; unfortunately, most traits are not controlled by major genes and molecular markers have had a limited success in this area [25]. Recently, simple molecular markers consisting in a single nucleotide substitution in the DNA chain (SNP) have been made easy and cheap to detect. This permits to use several thousand markers in each individual, thus all genes controlling a trait can potentially be associated to SNPs [26]. A main problem of this procedure is that these associations between SNPs and genes are lost after few generations of selection [27], but new associations can be reestimated. Nowadays genomics is being examined as a promising tool for many genetic programs, particularly in species like dairy cattle in which there is a continuous recording, the trait is expressed only in females, and generation intervals are large. In this case, genomics can be used for a better evaluation of young bulls that still have no offspring. Other uses of genomics will appear in the forthcoming years and it will be probably established as a useful complementary tool to current genetic programs.

### Animal Breeding and Sustainability

Animal breeding consists essentially in selecting animals kept in close reproduction systems, often

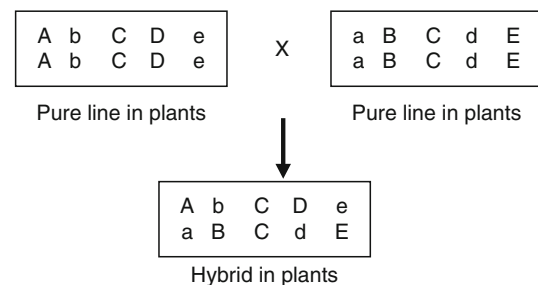
accompanied with crosses between these groups of animals [2]. Historically, the groups of animals kept in close reproduction were breeds, although modern intensive meat production of prolific species is now based in selection of synthetic lines. These lines are called “synthetic” because they do not correspond to traditional breeds, but have been generated by crossing animals from different breeds or crossing commercial “hybrids” (which are not hybrids in a genetic sense, as it will be seen later). This procedure allows obtaining a large genetic variability available for selection on productive traits. The relevance of breeds for sustainability lies in that some breeds can be particularly well adapted to local conditions, although this does not mean that local breeds are always better for local conditions than foreign breeds. A foreign breed can be better adapted or can be economically more interesting than a local breed. This is common mainly in species like poultry, pigs, or rabbits, which are usually kept in better environmental conditions than beef, sheep, or goats, but it also happens in ruminants. For example, Nelore cattle, a foreign breed in Brazil, has had a high success and now it is extensively implanted there [28].

### Definition of Breed

There is no consensus about the definition of what a breed is. Many definitions of breed have been compared [29–31], and the only common requirement to all of them is the genetic homogeneity, which applies essentially to external traits. It can be said that *a breed is a group of animals with some common external characteristics defined by some people who consider this group of animals to be a breed*. A breed requires some people who decide the external characteristics used to define the breed; often they also attribute some “average performances” to the breed. The problem with this definition is that it depends too much on external characteristics that may be very useful for dog or ornamental animals, but not necessarily for animals producing meat or milk in an efficient way. Some breeds were historically selected for improving some traits and they have been established as the most productive ones in intensive production systems; Leghorn hens for white eggs, Friesian cows for dairy cattle and Landrace and Large white in pig production are now widely

established. However, the word “Leghorn” or “Landrace” only define the external appearance of the breeds; there are many types of Landrace in the world, depending on the traits for which they have been selected, and the few multinational companies that control the eggs market use specific highly productive Leghorn lines, therefore the concept of “breed” is often of little utility. Other words used in animal breeding that can lead to confusion are “pure breed” and “hybrid.” In plants, a hybrid is the cross of two pure lines. A pure line is homozygous for all its genes, and all individuals have the same genotype, all hybrids have also the same genetic composition, and the cross of two plant hybrids produces very different plants due to the segregation of all the alleles (Fig. 1).

There are no “pure lines” in animals in the same sense as in plants. Pure lines in plants have been produced by self-fertilization or by fertilization of close relatives, something that is not possible in animals. Some attempts of creating highly inbred lines in pigs and poultry were done in the 1940s and 1950s, without positive results, because inbreeding produces infertility and abnormalities to a degree that prevents its use in animal breeding [2]. “Pure lines” in animals are only groups of animals in closed reproduction that will not be homozygous for all their genes, therefore animal “hybrids” will be crosses between lines or breeds with no genetic homogeneity. Moreover, it is a frequent practice in animal breeding to open the lines to some animals from other commercial lines in order to reduce inbreeding. This practice is also useful to capture genes that would be in lower frequency in the recipient line



**Animal Breeding Methods and Sustainability. Figure 1** Pure lines and hybrids in plants. Couples of letters indicate genes; capital letters indicate one allele of a gene and small letters another allele of the same gene

and that may be in higher frequency in the imported animals [32]. As “animal hybrids” are only crossed animals, they can be used to produce new “animal pure lines” with high genetic variability available for selection; for example, several rabbit breeds used for commercial purposes were originated by crossing commercial “hybrids” [33].

Breeds were created by humans after domestication by selecting traits they particularly liked. New breeds can be created nowadays. Apart from pets, many companies of pigs, rabbits, and poultry now use synthetic breeds without giving special importance to external characteristics, with the exception of the functional ones.

### Breed Conservation

Some breeds, local or not, can perform better than some intensively selected lines in systems in which food is less rich in protein or energy, or less balanced than in intensive systems. Some breeds can also perform better in some areas in which climate or breeding conditions are very different from the ones of current intensive production systems. There are more reasons for conserving breeds [34]: keeping genes that may be useful in the future, supporting sustainable animal production systems for food security, maintaining genetic variability for further use, conserving cultural heritage, etc. However, when a breed is useful, it does not normally need special aids for conservation, since it produces some profit and then it is kept for obtaining benefits. Help is needed especially for breeds that are not profitable, but there are reasons for inferring that they have genes that may be useful in the future. A question then would be whether the object of conservation should be breeds or genes, that is, whether it can be created as synthetic breeds having the genes of interest instead of spending funds in several programs for conserving several breeds. Although focusing the problem in keeping genes seems to be simpler, this can produce some problems. A first problem is that creating synthetic breeds may lead to undesirable gene interactions, difficult to manage for both the survival of the breed and the transmission of the interesting genes. Another problem would be the difficulties in integrating new synthetic breeds in areas in which farmers would not be prepared or accustomed to manage [35].

One of the main objectives of breed conservation, keeping genes for the future, has been discussed [35]. This objective is too vague unless the concrete purpose for using these genes in the future is envisaged. When a breed is a tool for making meat, milk, or eggs, conservation should be focused on whether this tool works now or whether there are expectations for using this tool in the future. This is an important point, because the extinction of a breed is completely different from the extinction of species. Breeds extinction, which can be created, transformed, or recovered, should not be compared with losing unrecoverable species created by natural evolution and forming part of a peculiar ecosystem.

The more concrete objective of maintaining genetic variability can be attractive for two reasons. First, genetic variability is needed for selection. Second, genetic variability implies a gene reserve that may also be useful when a rapid change in selection objectives is needed, for example, the current fertility problem of Holstein, partially caused by the increasing levels of inbreeding, can be managed by crossing Holstein with more fertile breeds [36, 37]. We should, however, notice that the genes of interest in animal breeding control economically relevant traits, thus keeping genetic variability is not an objective if the trait is near its optimum (100% of survival, for example). Genetic variability can be divided in between breeds variability and variability within breeds. It is important to know how much of the total existing genetic variability can be found between and within breeds, because if most of the genetic variability is contained within breeds, there is no genetic reason for conserving many breeds. For example, measuring the number of SNPs per kb in chicken, the International Chicken Genome Sequencing Consortium [38] detected “surprisingly little difference in diversity in comparisons between red jungle fowl and domestic lines, between different domestic lines, and within domestic lines.” For productive traits, it is generally admitted that about 50% of genetic variability is between breeds and 50% within breeds [39, 40]. Some methods of measuring genetic variability, like estimating genetic distances between breeds by molecular markers, have among other problems that they do not consider within breed genetic variability. The core of the argument for maintaining between breeds genetic variability is that

some breeds have genes that other breeds do not have or have in low frequency, and these genes may be useful in the future. It is a type of “insurance argument”: insurance against changes in market or environmental conditions, and safeguard against potential emerging disasters as emergent diseases [4]. There is nothing wrong in keeping every breed in danger when having an unlimited amount of financial resources, but when resources are scarce, for example, in developing countries, a precise analysis of the foreseen benefits is needed.

### Breeds and Sustainable Systems

By animal breeding sustainable systems, it is generally understood that farming systems are capable of maintaining their productivity indefinitely without damaging the environment [41]. This definition does not prevent having intensive systems with highly productive animals integrated in an industrial food chain, but sustainability is often associated to some kind of traditional farming at small scale in which waste is recycled, local breeds and local sources of food used and a rather high amount of hand labor is needed. Local breeds have a key role in this second type of sustainable systems, particularly when the environmental conditions are harsh or the food resources are not particularly good. This second type of sustainable systems is in general much less efficient for producing meat or animal products than intensive systems. There are, however, some reasons for establishing them:

1. *There are harsh environments in which no other systems will work properly.* A common example is cattle in swamp tropical areas. This applies essentially to cattle, sheep, and goats, and not necessarily to pigs, rabbits, or poultry, which have been kept in much better conditions traditionally.
2. *Using these systems in poor areas avoids land abandoning and migration of people to urban areas, avoiding desertification.* If life in these areas is hard for humans, this type of sustainable system should be considered as a temporary solution, because people living there deserve a better life.
3. *Sustainable systems are more environmentally friendly and produce a better animal welfare.* Although this reason is frequently invoked, this may or may not happen, and each case should be critically examined. Intensive industrial egg production can use enriched cages and manure process ensuring both welfare and sustainability. Moreover, animals in intensive systems arrive to commercial slaughter weight much earlier, thus they can produce less CO<sub>2</sub> and pollutants *per unit of product* than animals bred in extensive production systems, including pollution producing for transport, machinery, etc. A report ordered by the British government to the University of Cranfield [42] shows how this happens in poultry meat production, being organic chickens more contaminant *per kg of meat produced*, although results are more variable in pig production (for most pollutants, organic pigs contaminate less per kg of product). The same can be said about welfare: free-range hens are not necessarily happier than hens in enriched cages [43]. Looking for better animal welfare is not a particular task of industrial systems; it affects non-intensive systems as well.
4. *Some of these systems provide farmers an independence from big multinational companies.* This may be true, but is not necessarily good. Feeding people is a priority of poor countries, and the cheapest way may be to buy the genes to multinational companies. Genetics is very cheap; the genetic cost of 1 kg of pork, chicken, or rabbit meat is less than a 1% of the total cost of the meat as it will be seen in next section, and the same can be said about the genetics of 1 l of milk. Few companies provide the cheapest animal protein in the world (eggs and poultry meat and, up to a certain extent, pork meat), and genetics of dairy cattle is now managed in what is a world nucleus in practice. Poor countries need efficient genetic material for meat production even if this does not ensure genetic independence from multinational companies; this happens in industrial products and in other sectors (cars, industrial products, energy, etc.), and there is no reason for not accepting this in animal breeding.
5. *Some breeds are better adapted to local environment.* As said before, some breeds can be particularly well adapted to local conditions, although this does not mean that local breeds are better for local conditions than foreign breeds. There are spectacular examples of foreign breeds particularly well

adapted, as Nelore cattle in Brazil. Besides, adaptation is a bigger problem in some species than in others. Poultry, pigs, and rabbits have been raised in better environments than sheep or goats, thus intensive commercial breeds have less adaptation problems than in other species. Local food sources are often of lower quality than the usual food provided for highly productive breeds, and it has been said that local breeds can take a better profit of it. This is highly speculative, since the available information for these local breeds is normally scarce or null. Moreover, highly productive breeds of pigs, poultry, or rabbits can be bred with success in developing countries, even by small farmers [44]. Table 1 shows that small farms in rural conditions can obtain a similar profit as better farms using the same genetic material of a big multinational company (PIC). Local breeds of cattle, sheep, and goats may be better adapted in some harsh environments, although it is important to check whether this is true and when it is true.

6. *Local breeds produce better quality products.* The question is too general to give a simple answer. It is rather obvious that an Iberian pig (local breed) produces a much better cured ham than a Large White pig. Production of high quality products is one among several reasons for keeping breeds that are less efficient in producing meat or meat products. It is nevertheless convenient to check whether this better quality is detectable by the consumer. Some products like fresh cheese are not easy to differentiate, and local breeds sometime only show an external appearance of the animals different from the main breeds used for cheese production. It is also important, as St. Clair Taylor has stressed many times [45], that comparisons between breeds are performed at the same stage of maturity. As breeds have often different adult size and growth rate, if they are slaughtered at the same

commercial weight, they can be compared at different stage of maturity, thus differences between them can be due to the fact that one breed is younger, in physiological terms, than the other. For example, a breed can have a better meat quality than another only because at the same commercial weight it is slaughtered at a more mature stage.

## Animal Breeding Methods and Schemes

### Breeding Companies: Organization and Diffusion of Genetic Progress

Animal breeding can be practiced at small scale by farmers or small farmers associations, but this affects only the local breeds and its efficiency is low [21]. Nowadays animal breeding is generally in the hands of multinational companies or large cooperatives, although there are still medium sized ones performing animal breeding at a smaller scale. There are two types of schemes, based on recording data on farm or concentrating on all animal improvement in a small nucleus and diffusing later the genetic progress. The first scheme applies mainly to dairy cattle, and the second one to pigs, poultry, and rabbits.

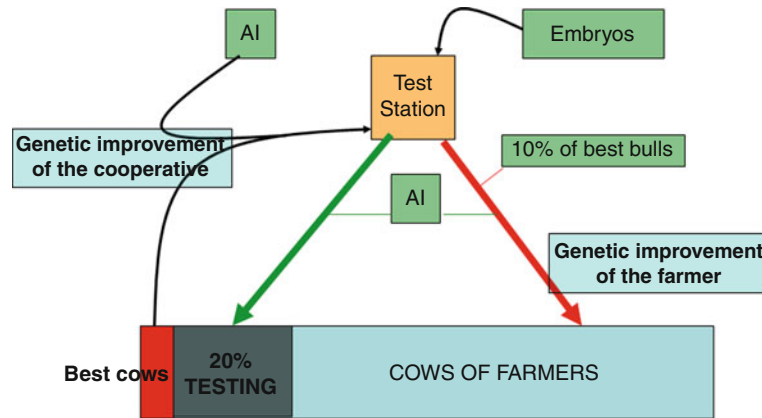
The standard example of the first scheme is dairy cattle. A 20% of the cows of a cooperative are inseminated with semen of young bulls that are going to be tested. The daughters are then inseminated with semen from other bulls in order to have lactation. Milk, protein, fat and cell count of the milk, and sometimes longevity, are recorded for each of the daughters, and these data are used to decide which 10% of the bulls being tested will pass to the catalogue of the cooperative (Fig. 2), to be used by the farmers to inseminate their cows in order to replace their stock. Each bull being tested provides semen for 1,000 cows in order to be sure that most of them will have at least 100 daughters, in order to achieve a high precision in the estimation of their breeding value [46].

**Animal Breeding Methods and Sustainability. Table 1** Sow reproductive performance of PIC pigs in Philippines [44]

|                   | Ave total born per litter | Ave born alive per litter | Ave weaned pigs per litter | Ave birth weight (kg) | Ave 30 day weight (kg) |
|-------------------|---------------------------|---------------------------|----------------------------|-----------------------|------------------------|
| Large farm sector | 10.7                      | 10.2                      | 9.2                        | 1.4                   | 7.7                    |
| Small farm sector | 11.9                      | 11.4                      | 11.1                       | 1.5                   | 8.8                    |

Source: From Gibson et al. [44]





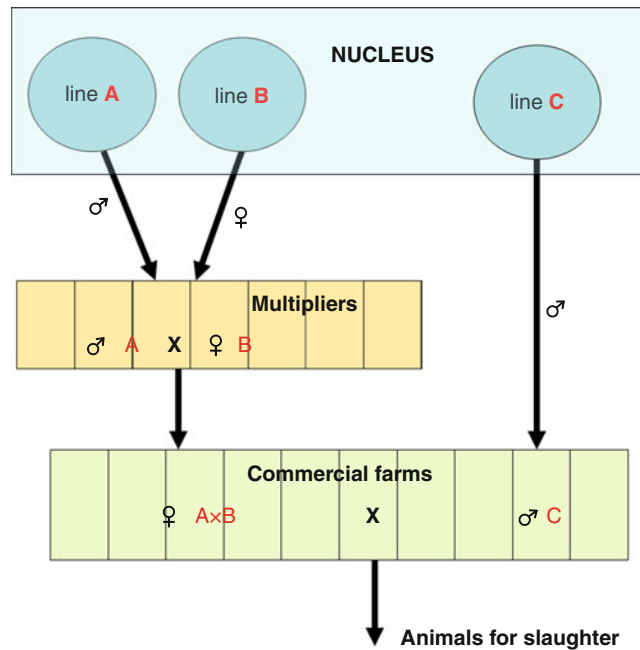
**Animal Breeding Methods and Sustainability. Figure 2**  
Schemes of genetic evaluation and gene diffusion for dairy cattle

This implies that an association created for bulls testing should have at least 100,000 cows in order to include a couple of bulls per year in their catalogue. Nowadays there are many practices: big cooperatives test their bulls; some associations test few bulls that are available in their catalogue after having 60 or 70 daughters, and import semen and embryos making them available to the members; some companies test bulls and then commercialize the semen; etc. In a global society in which frozen semen can be bought worldwide and records are collected in different countries, a global genetic evaluation has been established by an association called Interbull that publish their world evaluation for all sires of different countries.

Selection is made from the records, but since a bull being tested will inseminate 1,000 cows, a previous strong selection is made when deciding which bulls will go to the test station to be proved. To do this, the best cows of the association are inseminated with the best semen available to produce the bulls to be proven. Nowadays it is also possible to buy embryos from the best cows evaluated in the world and the best semen available. Genomics is used here to help in the evaluation of these bulls that will arrive to the station. A particularity of the system is that individual farmers can make their own genetic improvement. Catalogues contain an accurate prediction of the genetic value of bulls for many traits, thus a farmer having particular problems with protein content of the milk, functional conformation, or other trait, can buy semen from bulls particularly well evaluated for these traits, improving

the genetic level of his farm in the aspects he particularly needs.

The other scheme commonly used in animal breeding is the nucleus-multiplier scheme [47, 48]. Here all improvement is concentrated in a farm, from which it is spread to commercial farms through multiplication steps. This is the typical scheme for pigs, poultry, and rabbits (Fig. 3). Usually two lines are selected in closed reproduction, and males of one of the lines and females from the other are sent to farms called “multiplication units,” in which both are crossed to produce the crossed female sent to the farmers. Typically, these lines are selected for prolificacy and they may be selected for other traits. A third line is selected to produce the males that the farmers will use (called “terminal sires”); in this case, the line is not selected for prolificacy because this is a trait attributed to the dam, in which males seem to have little influence. Commonly, there is only one nucleus of selection in each company, and multipliers are spread in several countries. Multipliers act usually under a contract with the company; they buy parental stock for multiplication and they are in charge of providing facilities for breeding and commercializing the product: This system has allowed a rapid development of the business. There are some variations of the scheme; terminal sires are sometimes the product of a cross between two lines C and D, and sometimes there is a multiplication step more, in which other multipliers receive females  $A \times B$  to be crossed by a male from other line E to produce females  $(A \times B) \times E$  for the commercial farmers.



**Animal Breeding Methods and Sustainability. Figure 3**  
Scheme of selection and gene diffusion in pigs, poultry, and rabbits

Multiplication permits to reduce the cost of selection, for example, in pigs, a female coming from one of the lines of the nucleus and entering a multiplier can cost 600 € from which 500 € is the cost of the genetics and the rest is the cost of producing a pig. This female will produce about 15 crossbred females for production farms during her life, the rest of them being culled for various reasons (leg problems, diseases, etc.). This means that the 500 € of the genetic cost should be divided by the 15 females, giving 33 € of genetic cost for the farmer. If each female produces an average of 50 pigs for slaughter during his life, the cost of genetics for slaughtered pig is about 67 cents per pig, less than 1 cent for kg. These figures are extreme in poultry production, in which each female of the nucleus can produce nearly 100 females for the multiplication step, and each female of a multiplier can provide about the same quantity for commercial farms.

### Statistical Methods of Selection

Statistical methods used in animal breeding are essentially based in the infinitesimal model [49].

In this model, traits are determined by many genes independently distributed, having each one a small effect on the trait. A first consequence of the model is that genetically good animals can produce by chance some genetically poor sons, since by chance a son can inherit most of the alleles producing poor performances, whereas other sons can be genetically better than the parents if they get good versions of the alleles. As an average, all possible offspring of a parent will define how good this parent for breeding is. This is known as “breeding value” or additive value of the parent. The genetic value of an animal is not exactly this because genes can interact between them or among them producing better or worse individuals than the sum of their individual effects. These interactions are known as “dominance” when they appear between the two alleles of one gene or “epitasis” when they appear between alleles of different genes. Interactions can also occur between genotypes and environment, when the best genotypes in an environment (e.g., in the farm where the animals are selected) are not the best in other environments (e.g., in commercial farms). The development of artificial insemination in cattle and the prominent situation in the market of large

companies selling parent stock along the world has made the interactions between genotype and environment an important area of research in modern animal breeding [50].

Another consequence of the infinitesimal model is that it permits to invoke a theorem of statistics known as the *central limit theorem*, which permits considering the traits genetically distributed according to multivariate normal distributions. The multivariate normality has many advantages, for example, zero correlation implies independence between variables (which does not occur in other distributions), variables are determined by few parameters, and all relationships between variables are linear. Statistical methods in animal breeding are based thus in linear regression techniques. The most common models applied in animal breeding are called “mixed models” because they estimate simultaneously the breeding values, considered as random effects, and the environmental values, considered as fixed effects.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where  $\mathbf{y}$  is a vector with the data,  $\mathbf{b}$  is a vector containing the environmental effects (season, herd, parity, etc.),  $\mathbf{u}$  is a vector with the breeding values, and  $\mathbf{e}$  is a vector with the residuals.  $\mathbf{X}$  and  $\mathbf{Z}$  are known design matrixes containing 1s and 0s indicating the presence or absence of the effects. Fixed effects remain when repeating the experiment, and random effects change each repetition. Due to this, random effects are not usually estimated in classical statistical theory, but geneticists are interested in the value of these random effects, because they are the breeding values that, *as an average*, will be transmitted to the offspring; thus the best animals can be selected by taking offspring only from the ones with better predicted breeding values. The covariance structure of the breeding values is known due to our knowledge of Mendel’s rules for gene transmission. For example, half brothers share *as an average* half of the genetic information of their father. This allows calculating the genetic covariance matrix between random effects  $\mathbf{G}$  after knowing which part of the observed variance is due to the genes and which part to the environment. The most common method to estimate these variance components, correcting at the same time for the environmental effects, and using the same model as for

estimating breeding values, is called REML (Restricted or residual Maximum Likelihood) [20].

The data need not be normally distributed; in these cases, the model gives the best linear solution. Directly solving this model for many individuals, for example, several thousands or millions of data in dairy cattle, would not be possible, but an equivalent system of equations allows finding the solutions easily [18]. This system is known as Mixed Model equations, and the solution is known as the best linear unbiased prediction (BLUP) of the random genetic values.

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

There is a technical difficulty in solving the mixed model equations, because inverting a large matrix as  $\mathbf{G}$  is difficult. However, there is an easy way for directly calculating  $\mathbf{G}^{-1}$ , allowing a general use of these equations in animal breeding programs [19]. The model can be complicated adding repeated data, effects corresponding to single genes, and many other possibilities. It can also be used for many traits simultaneously. When several traits are used, the random effects are correlated not only due to the relationships between individuals, but also due to the genetic correlations between traits, originated when some genes have influence not in one trait but in several ones. Multitrait genetic variances and covariances can be estimated by REML as before, but Bayesian techniques, using a numerical procedure known as Markov Chain Monte Carlo (MCMC), have been particularly useful in complex situations, for example, when some traits have repeated data and other traits not and consequently the design matrixes  $\mathbf{X}$ ,  $\mathbf{Z}$  are not the same for both traits. Bayesian methods also permit to transform multivariate problems in series of unvaried estimations. Bayesian techniques with MCMC have been rapidly developed in the field of animal breeding, mainly for complex models, for example, when traits have different distributions, for censored data, for robust models, etc. (see [21] for a scope of their use and a comparison with classical methods and [23] for detailed description of Bayesian procedures).

In the case of using many traits, the objective is maximizing the economical benefit, which is obtained weighing each trait by economic weights. These weights can be calculated with more or less sophisticated

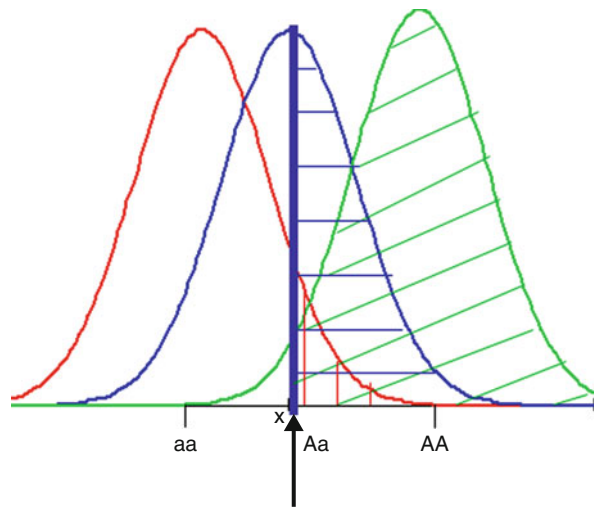
models [51, 52], but in essence they represent the amount of benefits, measured in economical units, obtained by improving one unity of the trait, for example, the number of euros of benefit for producing 1 kg of milk.

### The Use of Molecular Genetics in Animal Breeding

Molecular genetics has influenced modern genetic programs. Two different aspects, will be commented here, transgenic animals and molecular markers, including in the late genomic selection. A critical review and discussion about the uses of transgenesis and cloning in animal breeding, with references to markers, can be found in [24]. Genomic selection is very recent and its possibilities and development are still under discussion.

**Transgenesis** The first transgenic mice growing twice than normal created an enormous expectation about what could be done with transgenic animals [53], particularly in the field of animal production. However, few transgenic animals are now available, and the economical advantage of transgenic animals is small [24]. Although apparently it is economically viable to produce transgenic products useful for human health, the application of transgenic animals in medicine will not be considered here.

To apply transgenesis in animal production, genes with major effects are needed, but unfortunately, most economically interesting traits are determined by many genes of small effects. Sometimes there are genes with major effects for some traits, for example, for fat deposition in pigs, but classical selection has fixed yet the favorable alleles in commercial populations, thus they are not particularly useful now. When a trait of an economic interest has a major gene segregating in the population of study, this gene can be easily captured by selection. This can be shown by computer simulation [54, 55] but a simple example can help in understanding this. In Fig. 4, it can be seen the phenotypical distribution of a trait is controlled by a single gene. When selecting the best 50% of the animals, copies of the allele “A” are selected with preference. Therefore, in few generations of selection the gene will be in high frequencies or will get fixed. If the frequency of the favorable allele is low, the process takes more



**Animal Breeding Methods and Sustainability. Figure 4** Phenotypic distribution of a trait determined by a major gene with a high additive effect. Selection of the best 50% individuals

generations, but in general, it hardly will compensate to use transgenesis to capture it. Marker-assisted selection can be used for augmenting more rapidly the frequency of such genes of major effects, as it will be commented in next section.

Some major genes that are present in a breed or a line but not in other can be easily introduced by introgression without requiring transgenesis. The breed with the gene of interest *G* is crossed with the breed objective *O*, and then backcross is made by crossing *O* with the animals of the  $G \times O$  cross that carry the gene of interest. After several backcrosses, the gene is introgressed. An example of gene introgression is often performed with the Boorola gene in sheep that augments litter size, due to the high prolificacy of the carriers that permit an easy identification. When the carriers do not clearly show the gene of interest, genetic markers can be used to help the introgression [56].

The process of transgenesis is extremely inefficient. Genes are placed at random, thus the gene can be inserted in an inappropriate tissue or it can happen that genes around the inserted gene modify the expression of it. Transgenes are not always expressed and they are not always transmitted to descents. Moreover, many animals are needed for obtaining a viable embryo expressing the genes transferred. For example,

36,500 embryos were needed to obtain 18 transgenic calves expressing the trait, and the cost of each transgenic cow was higher than 500,000 dollars [57]. Lentivirus vectors can produce transgenic animals more efficiently in some species and at a lower cost, but they still suffer the former problems [58].

Transgenic animals should be tested to prove that they are commercially viable [24]. They should be tested for the trait that is the object of transgenesis, because it should be proved that the transgene is expressed in the animal and in the offspring for several generations. They should be also tested for commercial traits, since a transgenic line might be good for a trait but might have a poor productivity for other economic traits. The overall productivity should be evaluated. Transgenic animals may have poor fitness, sensitivity to diseases for which non-transgenic animals are resistant and poor performances in other traits that might affect longevity; it is also frequent that transgenic animals have reproductive problems.

Once the major gene has been transferred in an animal, a whole population or line having this gene has to be constituted. In the nucleus-multiplier scheme, inbreeding depression will increase when creating the transgenic nucleus, since mating with relatives during several generations are needed to spread the gene [59]. The process of evaluation of transgenic animals, and the diffusion of the transgene in a line, increases the genetic lag between the transgenic line and the commercial lines, due to the genetic improvement made during this time by its competitors. Diffusion of a transgenic animal in dairy cattle, in which a world nucleus of selection is much higher than in prolific species, and generation interval is large due to progeny testing (6 years), has also been studied. It has been calculated that in a population of 10,000,000 cows, three generations later after the introduction of the transgenic founder (18 years later), the presence of the gene in the population would be between 1% and 4% [60]. The genetic lag produced, the fact that a transgenic animal may be genetically inferior for other traits not controlled by the gene transferred, the complications of the processes and the scarce number of gene candidates for being transferred, makes transgenesis little attractive, even if it would be a less expensive and more successful technique [24].

**Genetic Markers and Genomic Selection** Genetic markers are parts of the DNA molecule that can be identified in individuals. They may be close to a gene of interest, so they can be used to select the favorable version of a gene affecting a quantitative trait. Genes controlling a quantitative trait are called QTL (quantitative trait loci), and occasionally they can have a large effect and can be selected with the help of a marker. However, generally quantitative traits are controlled by many genes with small effects, thus the effectiveness of markers has been rather limited [25]. The situation has dramatically changed since it has been possible to obtain a large number of markers at low cost and since they can be associated to many of the genes controlling traits even having small effects. There are several types of markers; the simplest one is the Single Nucleotide Polymorphism (SNP), which marks a place in the genome in which there is variability in a single nucleotide. Nowadays there are microchips allowing detection of about 50,000 SNPs in a genome; the number of SNPs that can be easily detected is increasing to 500,000 and soon it will be possible to genotype the whole genome of livestock species at reasonable prices. Prediction equations can be fitted, in which a set of SNPs will be used for predicting breeding values. Taking data from 1,000 to 4,000 animals (calling this *training population*), the model to be fitted can be

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_{50,000}x_{50,000}$$

where  $x_1, x_2, \dots, x_{50,000}$  are the variables indicating the presence of one polymorphism (AA, Aa, aa) of each SNP (usually indicated by 1, 0, -1, or by 0, 1, 2), and  $b_1, b_2, \dots, b_{50,000}$  are the regression coefficients to be estimated. These equations cannot be solved by least squares given the high number of SNPs in relation to the data available for the prediction, and Bayesian techniques should be used. The use of prior information allows solving these big equation systems, and depending on how prior information is included, the Bayesian methods differ [61]. This method can also be used for several traits [27]. Many of these SNPs are noninformative, and there are some techniques to select only informative SNPs [61, 62]. There is now a promising research area for selecting informative SNPs for prediction, often using nonparametric statistics [62].



Genomic selection has been proposed for traits that are difficult or expensive to measure (e.g., adult weight [24], index of conversion [63], mortality [64]). It has also been proposed in dairy cattle, in which the traits of interest are expressed in dams but selection acts mainly in sires, and the generation interval is very long [65]. It may be useful for other traits like litter size, difficult to select due to their low heritability, but studies are needed to determine its usefulness in these cases, because very low heritabilities will give poor prediction equations since the records will then be determined mainly by the environment.

A main problem of genomics is that the association of SNPs with the genes responsible of the trait quickly disappears in few generations of selection, thus the prediction equations have to be reestimated and new training populations are needed. Figure 5 shows an example of the loss in accuracy of the prediction of genetic values. It can be observed that accuracy is practically halved in four generations of selection.

This limits the use of genomic selection in current programs, because in some species the generation interval is short (6–9 months in rabbits or hens,

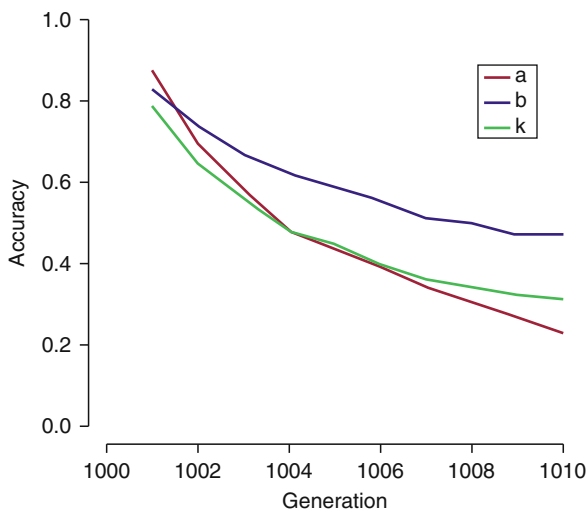
1 year in pigs), and a continuous reestimation can be difficult or expensive. Finding when and how genomics can be included in current genetic programs is one of the most important research areas nowadays.

## Future Directions

### The Future Evolution of Methods and Schemes

Prediction of breeding values from records seems to be well established with the methods briefly exposed in section “Statistical Methods of Selection” and it does not seem that dramatic changes will occur in the future at short or medium term. The revolution in methods for estimating breeding values is in the area of genomics. The possibility of having information from several thousands of markers at a reasonable price, now from several hundred thousands and in the near future from the whole genome, has brought the problem of how to manage all these data, and prediction methods are examined from other areas of knowledge as artificial intelligence, using nonparametric or semi-parametric methods, Bayesian methods, etc.

Schemes of selection are also changing due to the globalization of the market of genes. Today the best cows of the world are not dedicated to produce milk but embryos that are sexed, frozen, and commercialized. Some of the deficiencies of current dairy cattle programs such as long generation intervals can be partially solved by using genomic selection and having a quicker and better evaluation of the bulls being tested. Larry Schaeffer suggested that genomic evaluation can substitute progeny test, dramatically shortening generation intervals [66], but it is doubtful that farmers will accept genomics evaluation as they accept now tests mainly based in offspring records [65]. It can also happen that private companies will compete with others or that breeders can organize brands in which semen is not identified, like in pigs, as Maurice Bichard suggests [67], but it looks unlikely, since farmers like to perform their own genetic improvement at farm level by buying semen from accurately tested sires. Poultry genetics is now in the hands of two large holdings, and the only change envisaged in their structure is related to possible troubles with laws about competence. Pig companies tend also to be bigger, but they will probably coexist with nucleuses of smaller companies well



**Animal Breeding Methods and Sustainability.** Figure 5 Loss of accuracy of genomic selection on parameters a, b, and k of Gompertz growth curve. Selection acts on trait a, and the loss of accuracy of traits b and k are due to genetic correlations with trait a (From Blasco and Ibáñez-Escriche [27])

established in local markets, and with large pig production companies producing parental stock for themselves. Both pigs and poultry companies will introduce genomics in their programs not only as a complementary tool for selection but also as a commercial strategy, using modern methodologies as an added value to their products. The interest in meat quality traits and quality of animal products will probably increase. Companies will also stress the sustainability of their productions and the good welfare of their animals, thus there will probably be an increasing interest in traits like robustness and disease resistance. The new emphasis in sustainability will give importance to breed conservation programs, which will receive more attention and will get substantial public funds. Nevertheless, no dramatic changes in objectives are envisaged in the near future. Changes in genetic objectives are slow and the product of the selection arrives with delay to the market, thus this prevents short-term selection policies.

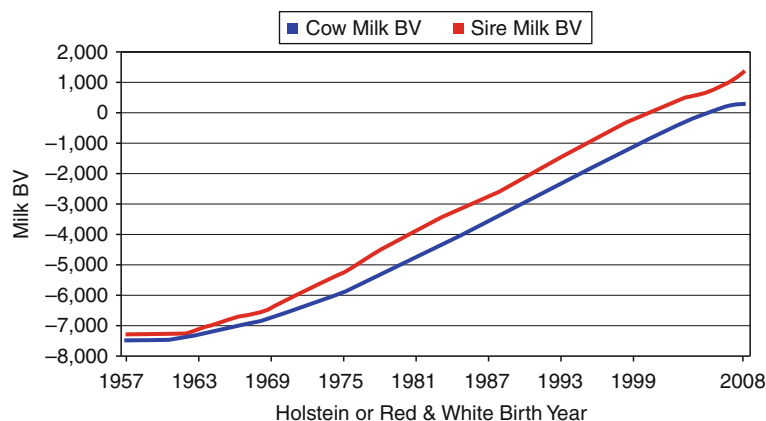
### The Limits to Genetic Progress

The theory of selection limits was developed by Alan Robertson (1920–1989) [68]. Classic quantitative genetics theory predicts the extinction of genetic variability by selection, and consequently the end of genetic progress. Frequencies of favorable genes increase with selection until they are close to 100%, and the genetic response is necessarily low, or genes are fixed by genetic drift, which occurs more likely when they are at high or

low frequencies and when the selected population is small. Mutation can introduce new genetic variability, but useful mutations are rare and they were disregarded in the classical theory of limits of selection. A decline in genetic response is thus expected until genetic variability is exhausted, and some experiments arrived to a plateau after showing response to selection along 20 or 30 generations in *drosophila* [69] and mice [70]. However, there is little evidence of any loss of genetic variability in commercial populations [71, 72]. Heritability of milk production in dairy cattle is not decreasing with time but augmenting! [73] and this is not only due to a better control of environmental variance or methods of correction, but also to the continuously maintained response to selection in the last 50 years (Fig. 6).

Long-term genetic responses have been observed in both plants and animals, and there are several examples of continuous genetic progress in all livestock species. A part of the success of the phenotypic trends observed in animals is due to improvements in nutrition, but when comparing chicken broilers fed with food as prepared in 1957 and as prepared in 2001, most of the observed differences are due to genetic improvement [74]. Figure 7 shows carcasses of poultry from an unselected line and a selected line of the same company, fed with modern food.

Broilers show a continuous growth, egg mass production continuously increases, pigs' lean growth selection has dramatically decreased the amount of fat of the carcasses, and in general all selection programs



Animal Breeding Methods and Sustainability. Figure 6

Genetic values of sires and cows in USA in the last 50 years (From USDA <http://aipl.arsusda.gov/eval/summary/trend.cfm>)

ACRBC Males - 2001 Feed



Ross Males - 2001 Feed



Day 43      Day 57      Day 71      Day 85

**Animal Breeding Methods and Sustainability. Figure 7**  
Carcasses of 1957 and 2001 of an unselected and a selected line of poultry fed with the same food [75]

continue having success [72]. The reasons for this apparent non-limits to selection are selection pressure on genes produced by mutation (which has a heritability of about 0.1% [72, 76]) or epistatic interactions, but even when epistatic interactions are important, additive variance typically accounts for over half, and often close to 100%, of the total genetic variance [77]. Bill Hill moved further the classical theory of limits of selection showing how new mutations with selective advantage can increase genetic variability [78]. An experiment corroborating the theory showed how totally homozygous lines produced artificially in *Drosophila melanogaster* could recover genetic variability by selection [79].

Are there limits to the genetic progress? Some traits have biological limits but still genetic progress can be obtained acting on related traits, for example, it is not possible to produce more than one egg per day, but it is possible to increase the laying period, and most of the new response to artificial selection in egg mass comes from this [72]. Highly productive animals can increase the incidence of pathological problems like ascites in broiler chicken or fertility in dairy cattle, but selection on these unfavorable traits [80] or crossbreeding [36] can be performed to continue the progress. Selection including traits different from strictly productive ones should be considered to avoid undesirable consequences of the continuous genetic progress [81]. Apart from some obvious limits (e.g., traits measured

in percentage, like survival, cannot surpass 100%), it seems that genetic response can be directed to overcome the biological limits presented when selection acts only in one or few productive traits.

## Bibliography

### Primary Literature

1. Wood J, Orel V (2001) Genetic prehistory in selective breeding. Oxford University Press, Oxford
2. Johansson I, Rendel J (1972) Genetics and animal breeding. Oliver & Boyd, Edinburgh
3. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443
4. Gandini G, Oldenbroek K (2007) Strategies for moving from conservation to utilisation. In: Oldenbroek K (ed) Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, pp 55–74
5. Darwin C (1859, 1964) On the origin of species (a facsimile of the first edition). Harvard University Press, Cambridge, MA
6. Trut LN (1999) Early canid domestication: the farm-fox experiment. *American Scientist* 87:160–169
7. Trut LN, Plyusnina IZ, Oskina IN (2004) An experiment on fox domestication and debatable issues of evolution of the dog. *Russ J Genet* 40:644–655
8. Achilli A, Olivieri A, Pellecchia M, Uboldi C, Colli L, Al-Zahery N, Accetturo M, Pala M, Kashani BH, Perego UA, Battaglia V, Fornarino S, Kalamati J, Houshmand M, Negrini R, Semino O, Richards M, Macaulay V, Ferretti L, Bandelt HJ, Ajmone-Marsan P, Torroni A (2008) Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr Biol* 18:R157–R158
9. Russell N (1986) Like enged'ring like. Cambridge University Press, Cambridge
10. Sageret A (1826) Considérations sur la production des hybrides des variantes et des variétés en général et sur celles de la famille des Cucurbitacées en particulier. *Ann Sci Nat* 8:294–314
11. Mendel JG (1866). Versuche über Pflanzenhybriden *Verhandlungen des naturforschenden Vereines in Brünn*, Bd. IV für das Jahr, 1865 *Abhandlungen*: 3–47. English translation at <http://www.mendelweb.org/Mendel.html>
12. Provine (2001) The origins of theoretical population genetics. The University of Chicago Press, Chicago
13. Fisher RA (1918) The correlations between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52:399–433
14. Lush JL (1937) Animal breeding plans. Collegiate Press, Ames
15. Smith F (1937) A discriminant function for plant selection. *Ann Eugen* 7:240–250
16. Hazel LN (1943) The genetic bases for constructing selection indexes. *Genetics* 28:476–490
17. Foote RH (2002) The history of artificial insemination: selected notes and notables. *J Anim Sci* 80:1–10

18. Henderson CR (1964) Selection index and expected genetic advance. In: Hanson WD, Robison HF (eds) Statistical genetics and plant breeding. NAS, NRC, Washington, pp 141–163
19. Henderson CR (1976) A rapid method for computing the inverse of a relationship matrix. *J Dairy Sci* 58:1727–1730
20. Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. *Biometrika* 58:545–554
21. Blasco A (2001) The Bayesian controversy in animal breeding. *J Anim Sci* 79:2023–2046
22. Gianola D, Fernando RL (1986) Bayesian methods in animal breeding theory. *J Anim Sci* 63:217–244
23. Sorensen D, Gianola D (2002) Likelihood Bayesian and MCMC methods in quantitative genetics. Springer, New York
24. Blasco A (2008) The role of genetic engineering in livestock production. *Livest Sci* 113:191–201
25. Dekkers JCM (2004) Commercial application of marker- and gene assisted selection in livestock: strategies and lessons. *J Anim Sci* 82:E313–E318, Suppl
26. Meuwissen THE, Hayes BH, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
27. Blasco A, Ibáñez-Escribano N (2010) Modifying growth curve parameters by multitrait genomic selection. 9th world congress on genetics applied to livestock production. Leipzig, 1–6 August 2010
28. Porter E (ed) (2002) Mason's world dictionary of livestock breeds, types and varieties. CABI Publ, Wallingford
29. Rodero E, Herrera M (2000) El concepto de raza. Un enfoque epistemológico. *Arch Zootecnia* 49:5–16
30. Hall SJG (2004) Livestock biodiversity. Blackwell, Oxford
31. Woolliams J, Toro MA (2007) What is genetic diversity? In: Oldenbroeck K (ed) Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, pp 55–74
32. Falconer D, Mackay TFC (1996) Introduction to quantitative genetics. Longman, Edinburgh
33. Baselga (2004) Genetic improvement of meat rabbits. Programmes and diffusion. In: Proceedings of the 8th world rabbit congress, Puebla, Mexico, pp 1–13, 7–10 September 2004. Available at <http://world-rabbit-science.com/WRSA-Proceedings/Congress-2004-Puebla/Puebla-2004-a.htm#Genetics>
34. Oldenbroeck K (ed) (2007) Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen
35. Blasco A (2008) Breeds in danger of extinction and biodiversity. *Rev Bras Zootecnia* 37:101–109, Special Supp
36. Madalena FE (2008) How sustainable are the breeding programs of the global main stream dairy breeds? The Latin-American situation. *Livest Res for Rural Develop* 20, article19. Available at <http://www.lrrd.org/lrrd20/2/mada20019.htm>
37. Hansen LB (2006) Monitoring the worldwide genetic supply for dairy cattle with emphasis on managing crossbreeding and inbreeding. 8th world congress on genetics applied to livestock production, Belo Horizonte, Brasil, 13–18 August 2006. Available at [http://www.wcgalp8.org.br/wcgalp8/articles/paper/1\\_484-1695.pdf](http://www.wcgalp8.org.br/wcgalp8/articles/paper/1_484-1695.pdf)
38. International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* 432:717–722
39. Hall SJG (2004) Livestock biodiversity. Blackwell, Oxford
40. Bennewitz J, Eding H, Ruane J, Simianer H (2007) Selection of breeds for conservation. In: Oldenbroeck K (ed) Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, pp 131–146
41. World Commission on Environment and Development (1987) Our common future. Oxford University Press, Oxford
42. Williams AG, Audsley E, Sandars DL (2006) Determining the environmental burdens and resource use in the production of agricultural and horticultural commodities. Main Report. Defra Research Project IS0205. Bedford: Cranfield University and Defra. Available at [http://randd.defra.gov.uk/Document.aspx?Document=IS0205\\_3959\\_FRP.doc](http://randd.defra.gov.uk/Document.aspx?Document=IS0205_3959_FRP.doc)
43. Webster J (2005) Animal welfare: limping towards Eden. Blackwell, Oxford
44. Gibson JP, Steane DE, Plastow GS, Warkup C (2006) Meeting the needs of rich and poor: the needs, opportunities and trends in pig genetic improvement. 8th world congress on genetics applied to livestock production, Belo Horizonte, Brasil, 13–18 August 2006. Available at [http://www.wcgalp8.org.br/wcgalp8/articles/paper/6\\_1040-2147.pdf](http://www.wcgalp8.org.br/wcgalp8/articles/paper/6_1040-2147.pdf)
45. Taylor CSt (1985) Use of genetic size-scaling in evaluation of animal growth. *J Anim Sci* 61:118–143
46. Goddard ME, Wiggins GR (1999) The genetics of dairy cattle. In: Fries R, Ruvinsky A (eds) The genetics of cattle. CABI Publ, Wallingford
47. Bichard M (1977) Economic efficiency of pig breeding schemes: a breeding company view. *Liv Prod Sci* 4:245–254
48. Hunton P (1990) Industrial breeding and selection. In: Crawford RD (ed) Poultry breeding and genetics. Elsevier, Amsterdam
49. Bulmer G (1980) The mathematical theory of quantitative genetics. Oxford University Press, Oxford
50. Knap PW, Wang L (2006) Robustness in pigs and what we can learn from other species. 8th world congress on genetics applied to livestock production, Belo Horizonte, Brasil, 13–18 August 2006. Available at [http://www.wcgalp8.org.br/wcgalp8/articles/paper/6\\_203-1825.pdf](http://www.wcgalp8.org.br/wcgalp8/articles/paper/6_203-1825.pdf)
51. Blasco (1995) Los pesos económicos en mejora genética animal. ITEA 91A:59–79
52. Weller I (1994) Economic aspect of animal breeding. Chapman & Hall, London
53. Palmiter RD, Brinster RL, Hammer RE, Trumbauer ME, Rosenfeld MG, Birnberg NC, Evans RM (1982) Dramatic growth of mice that develop from eggs microinjected with metallothionein-growth hormone fusion genes. *Nature* 300:611–615
54. Fournet F, Elsen JM, Barbieri ME, Manfredi E (1997) Effect of including major gene information in mass selection: a stochastic simulation in a small population. *Genet Sel Evol* 29:35–56

55. Villanueva B, Pong-Wong R, Grundy B, Wooliams JA (1999) Potential benefit from using an identified major gene in BLUP evaluation with truncation and optimal selection. *Genet Sel Evol* 31:115–133
56. Visscher PM, Haley CS (1995) Utilizing genetic markers in pig breeding programmes. *Anim Breedg Abstr* 63:1–8
57. Niemann H, Kues WA (2003) Application of transgenesis in livestock for agriculture and biomedicine. *Anim Reprod Sci* 79:291–317
58. Pfeifer A, Hofmann A, Kessler B, Wolf E (2004) Response to whitelaw: lentiviral transgenesis in livestock. *Trends Biotechnol* 22:159
59. Gama LT, Smith CS, Gibson JP (1992) Transgene effects, introgression strategies and testing schemes in pigs. *Anim Prod* 54:427–440
60. Hoeschele I (1990) Potential gain from insertion of major genes into dairy cattle. *J Dairy Sci* 73:2601–2618
61. Gianola D, De los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183:347–363
62. Gianola D, De los Campos G (2008) Inferring genetic values for quantitative traits non-parametrically. *Genet Res* 90:525–540
63. González-Reco O, Gianola D, Rosa GJM, Weigel KA, Kranis A (2009) Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genet Sel Evol* 41:3. Available at <http://www.gsejournal.org/content/41/1/3>
64. González-Reco O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendaño S (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* 178:2305–2313
65. König S, Simianer H, Willam A (2009) Economic evaluation of genomic breeding programs. *J Dairy Sci* 92:382–391
66. Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet* 123:218–223
67. Bichard M (2002) Genetic improvement in dairy cattle – an outsider's perspective. *Livest Prod Sci* 75:1–10
68. Robertson A (1960) A theory of limits in artificial selection. *Proc R Soc B* 153:243–249
69. Robertson FW (1955) Selection response and the properties of genetic variation. *Cold Spring Harb Sympos Quant Biol* 20:166–177
70. Roberts RC (1966) The limits to artificial selection for body weight in the mouse. II The genetic nature of the limits. *Genet Res* 8:361–375
71. Hill WG (2008) Estimation, effectiveness and opportunities of long term genetic improvement in animals and maize. *Lohmann Information* 43:3–20. Available at [http://www.lohmann-information.com/content/I\\_i\\_43\\_artikel1.pdf](http://www.lohmann-information.com/content/I_i_43_artikel1.pdf)
72. Hill WG, Bünger L (2004) Inferences on the genetics of quantitative traits from long-term selection in laboratory and farm animals. *Plant Breed Rev* 24, part 2:169–210
73. Van Tassell CP, Wiggans GR, Norman HD (1999) Method R estimates of heritability for milk, fat, and protein yields of United States dairy cattle. *J Dairy Sci* 82:2231–2237
74. Havenstein GB, Ferket PR, Qureshi MA (2003) Carcass composition and yield of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. *Poult Sci* 82:1509–1518
75. Havenstein GB (2006) Performance changes in poultry and livestock following 50 years of genetic selection. *Lohmann Information* 41:30–37. Available at [http://www.lohmann-information.com/content/I\\_i\\_41\\_2006-12\\_artikel5.pdf](http://www.lohmann-information.com/content/I_i_41_2006-12_artikel5.pdf)
76. Hill WG (1982) Predictions of response to artificial selection from new mutations. *Genet Res* 40:255–278
77. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* 4(2):e1000008. Available at <http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000008>
78. Hill WG (1982) Rates of change in quantitative traits from fixation of new mutations. *Proc Natl Acad Sci USA* 79:142–145
79. López M, López-Fanjul C (1993) Spontaneous mutation for a quantitative trait in *Drosophila melanogaster*. 1. Response to artificial selection. *Genet Res* 61:107–116
80. Pavlidis HO, Balog JM, Stamps LK, Hughes JD, Huff WE, Anthony NB (2007) Divergent selection for ascites incidence in chickens. *Poult Sci* 86:2517–2529
81. Rauw WM, Kanis E, Noordhuizen-Stassen EN, Grommers FJ (1998) Undesirable side effects of selection for high production efficiency in farm animals: a review. *Liv Prod Sci* 56:15–33

## Books and Reviews

- Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland
- Nicholas FR (2010) Introduction to veterinary genetics. Wiley-Blackwell, Oxford
- Willis MB (1998) Dalton's introduction to practical animal breeding. Blackwell, Oxford

---

## Animal Breeding, Long-Term Challenges

PEER BERG

Faculty of Agricultural Sciences, Aarhus University, Tjele, Denmark

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Sustainability  
 Sustainable Animal Breeding



Risks in Animal Breeding Schemes  
 Guidelines for Developing Sustainable Breeding  
 Schemes  
 Future Directions  
 Bibliography

## Glossary

**Animal breeding** Intentional breeding for certain traits, or combination of traits, by selecting animals for breeding with superior genotypes in growth rate, egg, meat, milk, or wool production, or in other desirable traits.

**Animal genetic resources** Genetic diversity, either characterized or as yet uncharacterized, that is found in economically important animals.

**Animal welfare** Animal welfare is the viewpoint that some or all animals, especially those under human care, should be treated in such a way that they do not suffer unnecessarily.

**Effective population size** Is the size of an idealized population that would behave the same as an actual population. The ideal population is one in which there is random mating and no selection. The effective size of a population is typically smaller than its actual size.

**Genetic trends** Changes in the mean breeding value of a population over time for one or more traits.

**Stakeholders** Any party that has an interest (“stake”) in a project or activity, which in the context of this chapter is animal breeding.

## Definition of the Subject

A major long-term challenge of animal breeding is to ensure that animal breeding is sustainable, in order to contribute to a stable long-term contribution of food for the globe. The need to address sustainability in animal breeding schemes has increased as the development in the last decades has been toward larger demands for food, fewer breeds contributing to the production of animal products, low effective population sizes despite the actual populations being large, and a decreasing number of breeding schemes providing the majority of the genetics underlying production of animal products.

The FAO report on *The State of the World’s Animal Genetic Resources for Food and Agriculture* [1] indicates

that the vast majority of developing countries have not been successful in sustaining genetic improvement in their livestock populations. Among the breeds considered to be in active use, 77% are located in developing countries. Animal genetic resources for food and agriculture (AnGR) provide the biological capital on which livestock production systems and food security are built. Planning for sustainable livestock development should, from the outset, take account of genetic differences among the species, the breeds, and the animals considered for use, along with their adaptive fitness to the production environments in which they will be kept. The different ways in which animals are used in different production systems and communities should also be recognized [2].

## Introduction

Animal production has significantly increased its productivity during the last 50 years. This has been a combined effect of improvements in the environment provided to the animals, for example, improved feeding and management, as well as efficient breeding schemes resulting in significant and cumulating genetic progress. These improvements are synergistic, as genetic improvements and management and feeding improvements stimulate each other. Examples of drastic changes in productivity include

- Manyfold increases in milk production of dairy cattle
- Improvements of growth, leanness, and litter size in pigs
- Manyfold increases of growth and egg production in broiler and layer hens, respectively

This trend in productivity is expected to continue for several reasons. In a competitive market for food production, genetic improvements are an essential parameter to ensure a profitable production. However, the demand for food is also expected to increase, and thus stimulates a larger and more efficient production. The World Bank has estimated that it will be necessary to increase meat production by about 80% between 2000 and 2030.

However, the improvements in productivity have also contributed to a global concentration on fewer breeds for the largest part of animal production.

In dairy cattle, Holstein-Friesian is now the dominating breed of the world, with exchange of semen and embryos across countries resulting in a relatively low effective population size ( $\sim 40\text{--}70$ , [3, 4]). In poultry breeding, a few multinational companies now control the commercial genetic improvement and thus the genetics of the vast majority of eggs and poultry meat produced globally.

Recently, the effect of animal breeding has come under critical evaluations from an ethical point of view, reflecting public concerns on the effects of current and future animal breeding. Developments in biology, neuroscience, and genetics have resulted in changes in our perception of the world, and there is an increasing focus on the effects of animal breeding and the ethics of changing animals by selective breeding for our own good (e.g., [5, 6]). Ethical problems in animal production have historically focused on housing and husbandry, but more emphasis is now focused on the effect of animal breeding [7]. It is argued that the fact that selective breeding can introduce welfare problems places an ethical responsibility on the animal breeding industry [7].

The European animal breeders have realized that these powerful technologies to create genetic progress have generated public concern about the impact of animal breeding and the use of new technologies [8].

According to FAO [2]: *A strategic and logistical approach to sustainable livestock development is required. To appropriately address the use of available AnGR and the role of genetic improvement in sustainable development, from the outset, all policies, plans, and programs for the livestock sector must:*

- *Be based on soundly established and agreed livestock development objectives (LDOs) and well-integrated and realistic livestock development strategies (LDS) that are able to achieve the LDOs.*
- *Account for major environmental, structural, and socioeconomic differences among the production systems concerned.*
- *Ensure participation of the end users (the livestock keepers themselves). Both men and women should have access to relevant information, be involved in the formulation of policies and plans, and have ample opportunities to give their opinions.*
- *Be appropriately funded.*

- *Promote step-by-step development and the sustainability of the actions undertaken.*
- *Be based on well-documented approaches that are understood and agreed by all the stakeholders involved at each stage.*
- *Take fully into account the fundamental principles of genetic improvement and their technical implications.*

Thus, a major future challenge of animal breeding is to ensure a sustainable use of the animal genetic resources available and to implement sustainable breeding schemes in active populations.

The focus of this entry is the long-term challenges of animal breeding focusing on sustainability of animal breeding schemes. This is done by discussing the concept of sustainability in the context of animal breeding and then describing risk factors affecting sustainability. Then guidelines for developing sustainable and future directions to ensure sustainable are discussed.

## Sustainability

Sustainability is a term that has a widespread use, and thus is often used in different meanings. It is often used in the sense of human sustainability on planet Earth and this has resulted in the most widely quoted definition of sustainability and sustainable development, that of the Brundtland Commission of the United Nations on March 20, 1987: “sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs” [9]. This definition is also relevant to animal breeding; it is, however, not very operational.

In order to define and address sustainability in the context of animal breeding, it is first necessary to identify the stakeholders of animal breeding schemes. These stakeholders all have interests in animal breeding, and some of their interests might be conflicting. Understanding the diverse interest in animal breeding is a prerequisite for developing sustainable breeding schemes.

Obvious stakeholders are the animal breeders, breeding companies, and breeding associations, which have the major responsibility for defining and running the breeding scheme. Additional stakeholders include producers, veterinarians, manufacturers and retail enterprises, and consumers. For some aspects, the society as a whole may be considered a stakeholder in itself.

These stakeholders have different perceptions of the desired outcome of animal breeding schemes and thus sustainability, influenced by their role in production and utilization of animal products and their time horizon.

In the short term, objectives of different stakeholders can be very different. Animal breeders will typically focus on the factors limiting the profitability of the primary production. The interest of manufacturers and retail enterprises is on product quality, quantity, and possibilities of adding value to the primary product to ensure profitability of manufacturing and retail. Consumers will typically focus on product quality and price, product safety in addition to potentially cultural and political issues, such as environmental impact and animal welfare.

However, it can be expected that stakeholders will share many long-term interests in animal breeding schemes. Long-term interests of producers and manufacturers are likely to be more similar to the interests of consumers in the long-term, as they are the main actors determining quantity and quality demands. A thorough discussion of sustainability in the context of animal breeding is provided in [10], focusing on the need to address the interests of all stakeholders.

Interests of stakeholders include the following areas, which should be considered for defining sustainable animal breeding schemes for the future [11]:

- Food security, including
  - Quantity produced
  - Product quality
  - Food safety from both a biological and technical perspective
- Socioeconomic effects, including
  - Impacts of breeding schemes on rural economy, employment, and trade
  - Subsidies and their effects on objectives and methods used in animal breeding
  - Public (ethical) perception of the methods and technologies applied in animal breeding
- Environmental impact, including
  - The relationship between animal production and measures of environmental quality
  - Landscape management where animal production plays a major role
  - Interactions with biodiversity and ecosystems

- Health, welfare, and ethics including
  - Health management and biosecurity related to both the risk of zoonoses and contamination of products
  - Ethics of breeding and production

### Sustainable Animal Breeding

Animal breeding is based on selecting animals with a superior genotype relative to the desired direction of selection, as reflected in the breeding objective. However, breeding values cannot be observed directly in animals, and they have to be predicted based on phenotypic records on the individuals themselves and their relatives and/or genetic markers. Thus, animal breeding is based on selecting the seemingly superior individuals given their predicted breeding values. Secondly, selected parents pass half of their genes to a given offspring, and each individual is thus a genetically unique individual. A direct result of this is that planning of animal breeding schemes is not an exact science. Genetic progress can be predicted but the realized response to selection can vary depending on a number of factors reflecting biological and economic uncertainty that influences the outcome of a breeding scheme. The breeding objective represents the desired direction of change of a population, most often expressed as the marginal economic value of changing a trait. The breeding objective should address all aspects of animal production and the characteristics of the product.

Ignoring the interests of some stakeholders when defining breeding objectives might negatively affect the success of a breeding scheme. This can be addressed by considering the interests of all stakeholders. The large potential of animal breeding is the contribution of additive improvements toward a long-term objective. However, there are factors that are not necessarily stable over time, factors that can potentially also change the interests of stakeholders. Examples of such factors include market structure, fluctuations in prices, regulations of production systems, disease outbreaks, and variation in the response to selection, both in the traits directly selected for and in traits indirectly affected as a correlated effect of selection. In addition to these factors, uncertainty on the assumptions underlying the design of a breeding scheme contributes to the

risk of this breeding scheme. The risk of a breeding scheme is related to the deviations of the realized effects of a breeding scheme relative to those planned. These factors have two effects: some contribute to a reduction of the realized outcomes of a breeding scheme (e.g., diseases) and others contribute to an increased variability of the outcome of a breeding scheme.

Thus, a sustainable animal breeding scheme should address the interests of all stakeholders and actively reduce the risks of the breeding scheme. This is in line with the SEFABAR project that produced a “Code of Good Practise for Farm Animal Breeding and Reproduction Organisations” [10, 12]. The code addresses issues of food safety and public health, product quality, genetic diversity, efficiency, environmental impact, animal health, animal welfare, and breeding and reproduction technologies. The code is based on the following six general statements:

- Breeding organizations must follow zootechnical, animal welfare, and animal health legislations and relevant regulations and practices.
- Breeding organizations must consult and collaborate with international, national, and regional authorities for the development and implementation of policies, practices, and regulations. These policies should assist the achievement of economic, environmental, and social sustainability of the animal breeding sector.
- Breeding organizations must use modern biosecurity methods to minimize disease transmission.
- Breeding organizations must ensure the health and welfare of the animals under their care.
- Breeding organizations must treat the animals under their care with respect.
- Breeding organizations must ensure that selection for production traits is balanced by appropriate attention to reproduction traits and health- and welfare-related traits.

The intention is that this should result in sustainable breeding programs by an economically viable balance of (a) food safety and public health, (b) product quality, (c) genetic diversity, (d) efficiency, (e) environmental impact, and (f) animal health and welfare.

## Risks in Animal Breeding Schemes

The expected outcome of a breeding scheme might not be the same as the realized outcome for several reasons. These include

- Uncertainty on the genetic model assumed
- Uncertainty on the assumed genetic parameters (heritabilities, genetic correlations, etc.)
- Genetic drift resulting from finite population size
- Suboptimal decisions and implementation of the breeding scheme
- Breeding objectives that do not include all relevant costs and benefits of the production system targeted
- Changes in markets and consumer preferences

Some of these can contribute to both positive and negative deviations from the expected outcome, for example, deviations in genetic parameters from those assumed might either increase or decrease genetic gain relative to that expected.

However, most deviations are expected to result in a realized outcome of selection being lower than that expected, as the predictions assume that parameters are known and selection decisions are optimal. If the breeding objective does not reflect the true breeding objective, this will always result in selection decisions being suboptimal and response to selection being lower than expected. Likewise, changes in markets, consumer preferences, or political regulations of the industry will result in suboptimal goals being targeted by the breeding scheme, and thus the realized outcome being inferior to that expected.

Thus, two groups of factors contributing to the risk of animal breeding schemes can be identified.

The first group is risks intrinsic to the breeding scheme. The response of a breeding scheme might deviate from expectations due to variability of response to selection for the traits considered in the breeding objective and in the selection index. This can be due to uncertainties about genetic parameters used for predicting breeding values, uncertainties in the breeding values predicted from phenotypic records and/or genetic markers, and genetic drift.

Genetic drift can also increase the frequency of deleterious (recessive) alleles. Examples include BLAD [13] and CVM [14] in dairy cattle and the halothane sensitivity locus in pigs [15]. These alleles increased in

frequency either by directly having a positive impact on some of the traits selected for, or being linked to alleles with positive effects on the traits selected for. In general, all traits not directly selected for and not correlated with the traits in the selection criterion are most affected by genetic drift. In a closed population, genetic drift is highly related to the rate of inbreeding.

Unfavorable correlated responses to selection can appear if important traits are not included in the breeding objective and selection criterion. There is evidence of general negative effects of selection for increased production resulting in increased occurrence of behavioral, physiological, and immunological disorders [16]. A theory (resource allocation theory) has been proposed to explain the general negative correlated changes in fitness and health-related traits when selecting for production traits. The theory predicts that with limited resources available, selection for increased production results in less resources being allocated to, for example, health and reproductive traits [17]. In a mice model it was shown that a line selected for large litter size allocated more resources to lactation and mobilized body energy for a longer time period than a control line [18]. However, results suggest that not all aspects of immune response are decreased by selection for production. In a comparison of chicken lines from 1957 to 2001, it was concluded that the 2001 chickens selected for growth have a decreased adaptive immune response but a better cell-mediated immune response [19]. These examples highlight the need to consider the direct response to selection not only in the traits selected for, but also in other traits of relevance for production, animal health, and animal welfare.

Effective population size (which in a closed population is inversely proportional to rate of inbreeding) is the most important indicator of genetic risk in a breeding scheme, being inversely proportional to both genetic drift in a closed population and variation in selection response. Larger effective population size results in less genetic drift and less variability in the response to selection. Selection intensity, population size, and selection and mating criteria all influence the effective population size. It has been argued that the effective population size should be at least 50–100 individuals in order to avoid negative effects of genetic drift [20], but should be significantly larger in order to maintain mutations contributing to maintaining genetic variability.

The second group of risks is external to the breeding scheme. These include environmental risks due to changes in environmental factors unforeseen when planning the breeding scheme. Such risks are diseases, particularly epidemic diseases and diseases under national and international regulations (e.g., foot-and-mouth disease), and changed market and/or production regulations. These changes will most often result in the realized outcome of a breeding scheme being less than expected.

### **Guidelines for Developing Sustainable Breeding Schemes**

Animal breeding schemes produce additive genetic improvement, and are thus most powerful when applied for a long-term improvement of a population. Thus, breeding objectives should reflect the future market for which production is aimed. The development of breeding objectives includes an assessment of the breed and its characteristics relative to the future market preferences. Traditionally, this includes an economic analysis aiming at estimating the marginal value of improving a trait. These marginal values are often the economic value of changing a trait given all other traits are constant. The breeding objective is then defined by the marginal values of the traits. However, this is not necessarily sufficient as a breeding objective should also consider the risk and uncertainty of aiming at a particular market, including status and trends in consumer preferences, social attitude, political and economic regulations, and infrastructure.

The breeding objective should reflect the most profitable improvement of the population, conditional on the current and future market conditions. The derivation should include traits beyond those currently recorded or included in a selection criterion.

The expected genetic trends should also be evaluated to ensure the breeding objective is sustainable. Ethical evaluations and public perception of animal welfare might alter the relative weights used in the breeding objective. Including a trait in the breeding objective does not guarantee that that particular trait is improved. Traits with low heritability, traits negatively correlated to important traits in the breeding objective, and traits measured late in life are at risk of not being improved despite being in the breeding objective. Traits that fulfill



all the above-mentioned criteria include health-related traits, such as mastitis and fertility in dairy cattle. It has been suggested that such traits should have a larger weight in the breeding objective, the larger weight reflecting a nonmarket value [21]. Such nonmarket values could reflect public concern on, for example, increased disease susceptibility resulting from the current breeding objective.

Ignoring important traits in deriving marginal (economic) values to be used in the breeding objective will bias the prediction of the economic outcome of the breeding scheme. For example, ignoring quality measures that are not considered in current market conditions but are important in distinguishing the product from other competing products might be harmful to the long-term competitive ability of the product.

Breeding objectives should not only be defined in terms of production, but should also include traits related to animal health and welfare resulting in robustness being part of the breeding objective [16].

Derivation of the marginal value of all important traits is crucial, in order to predict the effects of selection, not only on traits selected on, but also traits correlated to those traits. This is necessary to avoid unfavorable genetic changes in traits not selected for.

The expected outcome for a given breeding objective also depends on the recording scheme. Recording is often one of the most costly elements of a breeding scheme and should thus be an integral part of planning a breeding scheme. Generally, the objective of recording should be to allow for favorable genetic trends in the traits included in the breeding objective. This might be complicated by biological constraints of, for example, when a trait is realized and the number of individuals it can be recorded on. Examples are traits measured late in life (e.g., longevity), measured after slaughter (e.g., meat quality), or measured on one sex (e.g., female fertility), which poses challenges in defining efficient recording schemes. Modern technology allows for some remedies to this challenge. Genetic markers can be measured early in life, and in addition they can be used to predict the breeding value for traits not yet recorded [22].

In addition to the traits directly selected for, other important traits should also be recorded, in order to

document the trend in these traits. Such traits include traits that are economically or ethically important, but they are not necessarily included in the breeding objective if they are at acceptable levels. However, correlated effects of selection might change these traits in undesired directions.

Two aspects of sensitivity to environmental factors should be considered. Chance events, such as disease epidemics have the potential to critically damage a breeding scheme, particularly if breeding animals are kept in a limited region or in a few herds. The foot-and-mouth disease outbreak is an example of a disease outbreak that has threatened breeding nuclei. Backup and safety procedures are necessary for safeguarding a breeding scheme. This can be done by implementation of techniques such as cryopreservation of embryos, eggs and semen, and segmentation of the breeding nucleus. This will naturally be very species specific, depending on the biological possibilities in the given species.

The second aspect of environmental sensitivity relates to environmental sensitivity, whether an individual's performance is sensitive to the environment in which it is kept. More specifically, this is an issue if genotypes and environment interact, such that an individual would be ranked differently in two environments relative to other individuals. A more realistic example is offspring from two bulls having different relative merits in two environments. The implications of an interaction are multiple. First, prediction of the performance in an environment that differs from the environment in which selection was performed might be impossible. Second, genetic improvement in one environment might not result in similar genetic changes in another environment. As an example, in some breeding schemes, candidates for selection are housed individually in order to allow for individual recording of the traits of interest, whereas animals in commercial production are housed in groups, providing a significantly different environment to that selected in. In such cases, where the environment of the breeding nucleus and the production herds differ, then genetic trend should be recorded in the production environment to ensure that selection is in the direction desired. If genotype–environment interaction exists, genetic evaluations should take this into account. Also the breeding objective should be specific

on which environment(s) genetic changes are optimized for. A specific type of genotype–environment interaction is the situation where environments describe a continuum of environmental effects, and thus cannot be grouped in classes of environmental effects. Reaction norm models describe such interactions, and a feature of these models is that they can model the sensitivity to the environment, and that this sensitivity is partly genetically determined. This has two implications. First, it is expected that environmental sensitivity will change as a correlated effect of selection on the mean, assuming nonzero correlations between mean performance and sensitivity. Second, it implies that environmental sensitivity can be selected for or against. The second implication also means that genetic changes will require changes in the environment to be fully expressed in all environments. This is an important outcome of a breeding scheme, and it has to be taken into consideration whether such environmental changes can be accommodated in commercial environments. There are many examples of highly improved breeds, selected in environments with a high level of management, that fail to express their genetic potential when transferred to environments characterized by lower management levels, for example, when highly selected breeds are transferred to tropical countries with lower input levels and other environmental characteristics. Thus, potential changes in environmental sensitivity have to be taken into account when predicting and evaluating the expected response to selection.

The optimum selection environment is not always equal to the environment in which the response is to be realized, but depends on the degree of genotype–environment interaction (determined by the ratio of variances in slope and level of a linear reaction norm), the correlation between level and slope, and the heritability of the trait [23].

In relation to breeding objectives, it was stressed that the market for which production is aiming should be defined. The reasoning is that this is a prerequisite to ensure that consumers will accept the product and its characteristics and that the product(s) addresses the preferences of the consumers. Consumer responses have been instrumental in changing egg production systems in European countries, and the use of gene transfer worldwide.

Measuring the success of a breeding scheme involves two components. First, expected response to selection should be predicted, as the measure against which to compare the operational and strategic outcome of the breeding scheme. To predict the expected response to selection estimates of genetic parameters, covariances between traits, generation intervals, recording scheme, and resulting accuracies of the predicted breeding values, selection intensities and the weights used in the breeding objective are required. Uncertainty about one or more of these should be taken into account when predicting response to selection. These predictions are based on a genetic model, on which there may also be uncertainty. Second, realized genetic trends should be compared to the predicted response to selection, both in order to evaluate the success of the breeding scheme, but also to modify it if required to better fulfill the overall objectives of the breeding scheme.

Most breeding schemes rely on genetic progress accumulating in a nucleus, representing the animals active in the breeding scheme. The genetic progress is then disseminated to production herds. This might involve several tiers, such as a breeding and multiplier level, in order to efficiently disseminate genetically superior individuals or semen to production herds. The time from genetic progress in the breeding nucleus to the dissemination of this progress in production herds is termed genetic lag, and is a major indicator of the efficiency of utilizing genetic progress in production.

The outcome of a breeding scheme is realized in the years following the selection decisions and the majority of the costs. An economic evaluation of a breeding scheme thus requires that costs and revenues are discounted back to a common base. This should take into account that returns more distant in time are more uncertain than those occurring in the near future. This involves defining the time horizon for which revenues are discounted, reflecting the uncertainties about the market and production conditions. This reflects an economic time horizon of the breeding scheme.

Expected and realized response to selection should be compared at regular interval as an integral part of controlling and optimizing the breeding scheme. This should be done at intervals shorter than the economic time horizon.

## Future Directions

Following is a discussion including potential impacts on the development of certain areas of science.

Animal breeding is a competitive business, and producers pursue genetic material that best suits their needs. This naturally means purchasing animals that are economically optimal in the specific production. This potentially limits the focus on animal welfare-related traits not directly related to productivity.

However, since animal breeding decisions for most species are made centrally and, increasingly, internationally it has been argued that international agreements are needed [7] to ensure sustainable breeding schemes that address all relevant traits.

There is evidence that breeding companies are addressing these challenges. Survival in poultry has been improved by, among other initiatives, selective breeding and current focus is on selection against feather pecking, cannibalism, leg disorders, and heart/lung disorders [24]. And more generally, breeding objectives now contain more traits, also related to animal health and costs of production. In many cases, inclusion of health- and welfare-related traits results in more profitable breeding schemes than selection on production alone [25].

It should, however, also be acknowledged that despite the well-documented negative effects that selection for production traits has had on welfare-related traits, modern genetic technologies also give possibilities of understanding the biology underlying animal welfare and actively select for improved welfare [26]. Welfare-related traits are likely to show genetic variation, and it is thus also possible to select for increased animal welfare [27].

New technologies have the potential to affect animal welfare in both positive and negative directions [28]. Positive examples are the use of sexed semen in dairy cattle, which can decrease the number of unwanted male offspring that are killed after birth in some breeds and result in easier birth of female offspring. An overall positive effect is conditional on no adverse effects of the technology on the calves and/or the mothers.

Genomic selection [22] is currently under implementation in the large dairy, pig, and poultry breeding schemes. Genomic selection is fundamentally different

to previous selection strategies, which relied on phenotypic records being recorded on the candidates for selection and/or their relatives. With genomic selection, this is replaced by a two-step procedure where the association being genome wide, markers and phenotypes are established in a part of the population where both phenotypic recording and genotyping are performed. Secondly, breeding values can be predicted for individuals based on their genotype. This allows for accurate breeding values being available earlier in life. This is particularly useful for traits recorded late in life, traits with low heritability, and traits recorded in one sex, which are largely the characteristics of many traits related to animal health and welfare. Genomic selection thus improves efficiency of selection for traits that are currently difficult to improve genetically, and thus has the potential to contribute to more sustainable animal breeding schemes [29].

Effective population size is a measure of the risk of a breeding scheme. Constraining the rate of inbreeding has the potential to reduce the variance of response to selection, reduce the loss of genetic variance, and reduce genetic drift. Quadratic selection indices jointly maximizing response to selection while constraining rate of inbreeding have been developed [30]. This quadratic index optimizes the genetic contributions of parents to the next generation, and the contribution of any parent depends on which other parents that are selected.

Acceptable rates of inbreeding are often described as being in the range of 0.5–1% per generation, equivalent to an effective population size of 50–100. More specifically, the acceptable level of inbreeding depends on the extent of inbreeding depression on fitness [20], and thus a sustainable breeding scheme ought to validate the acceptable level of inbreeding.

The implementation of sustainable breeding schemes requires, as described above, a thorough analysis based on an understanding of genetics, economics, and market characteristics. Training and education is a prerequisite for qualified personnel to contribute to the development and maintenance of sustainable breeding schemes.

Sustainability is a new measure in the context of animal breeding, but many elements contributing to sustainability are well known and have been taken into account previously. The aim of developing sustainable

animal breeding schemes puts these elements into a coherent context, highlighting the need to take interests of all stakeholders into account, as well as the risks and uncertainties of assumptions and outcomes.

## Bibliography

### Primary Literature

1. FAO (2007) The state of the world's animal genetic resources for food and agriculture. In: Rischkowsky B, Pilling D (eds) Rome (Available at [www.fao.org/docrep/010/a1250e/a1250e00.htm](http://www.fao.org/docrep/010/a1250e/a1250e00.htm))
2. FAO (2010) Breeding strategies for sustainable management of animal genetic resources: FAO animal production and health guidelines, vol 3. FAO, Rome, 160 pp
3. Kim ES, Kirkpatrick BW (2009) Linkage disequilibrium in the North American Holstein population. *Anim Genet* 40:279–288. doi:10.1111/j.1365-2052.2008.01831.x
4. Sørensen AC, Sørensen MK, Berg P (2005) Inbreeding in Danish dairy cattle breeds. *J Dairy Sci* 88:1865–1872
5. Hastedt H (2000) Beats, humans and the ethics of animal breeding. *Arch Tierz* 43:218–225
6. Larrere R (2003) The issue of modern breeding a social request, consumer request or responsible inquiries? *Prod Anim* 16:329–332
7. Olsson IAS, Gamborg C, Sandoe P (2006) Taking ethics into account in farm animal breeding What can the breeding companies achieve? *J Agric Environ Ethics* 19:37–46
8. Neeteson-van Nieuwenhoven A, Merks J, Bagnato A, Liinamo A (2006) Sustainable transparent animal breeding and reproduction. *Livest Prod Sci* 103(3):282–291
9. United Nations General Assembly (1987) Report of the World Commission on Environment and Development: Our Common Future. Transmitted to the General Assembly as an Annex to document A/42/427 – Development and International Co-operation: Environment. Retrieved on: 2010-06-30
10. Gamborg C, Sandoe P (2005) Sustainability in farm animal breeding a review. *Livest Prod Sci* 92:221–231
11. Woolliams J, Berg P, Mäki-Tanila A, Meuwissen T, Fimland E (2005) Sustainable management of animal genetic resources. Nordic Gene Bank Farm Animals, Nordisk Genbank Husdyr. ISBN 92-893-1089-8
12. CODE-EFABAR (2010). <http://www.effab.org/CODEEFABAR.aspx>. Assessed 30-6-2010
13. Shuster DE, Kehrl ME, Ackermann MR, Gilbert RO (1992) Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. *Proc Natl Acad Sci USA* 89:9225–9229
14. Agerholm JS, Bendixen C, Andersen O, Arnbjerg J (2001) Complex vertebral malformation in Holstein calves. *J Vet Diagn Invest* 13:283–289
15. Carden AE, Hill WG, Webb AJ (1983) The inheritance of halothane susceptibility in pigs. *Genet Sel Evol* 15:65–82
16. Star L, Ellen E, Uitdehaag K, Brom FWA (2007) A plea to implement robustness into a breeding goal: poultry as an example. *J Agric Environ Ethics* 21:109–125
17. van der Waaij EH (2004) A resource allocation model describing consequences of artificial selection under metabolic stress. *J Anim Sci* 82:973–981
18. Rauw WM, Luiting P, Beilharz RG, Verstegen MWA, Vangen O (1999) Selection for litter size and its consequences for the allocation of feed resources a concept and its implications illustrated by mice selection experiments. *Livest Prod Sci* 60:329–342
19. Cheema MA, Qureshi MA, Havenstein GB (2005) A comparison of the immune response of a 2001 commercial broiler with a 1957 randombred broiler strain when fed representative 1957 and 2001 broiler diets. *Poult Sci* 82:1519–1529
20. Bijma P (2000) Long-term genetic contributions. Prediction of rates of inbreeding and genetic gain in selected populations. PhD thesis. ISBN 90-5808-241-5
21. Olesen I, Groen AF, Gjerde B (2000) Definition of animal breeding goals for sustainable production systems. *J Anim Sci* 78:570–582
22. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
23. Kolmodin R, Bijma P (2004) Response to mass selection when the genotype by environment interaction is modeled as a linear reaction norm. *Genet Sel Evol* 36:435–454
24. Flock DK, Laughlin KF, Bentley J (2005) Minimizing losses in poultry breeding and production: how breeding companies contribute to poultry welfare. *Worlds Poult Sci J* 61:227–237
25. Lawrence AB, Conington J, Simm G (2004) Breeding and animal welfare: practical and theoretical advantages of multi-trait selection. *Anim Welfare* 13:S191–S196
26. Jensen P, Buitenhuis B, Kjaer J, Zanella A, Mormede P, Pizzari T (2008) Genetics and genomics of animal behaviour and welfare – Challenges and possibilities. *Appl Anim Behav Sci* 113:383–403
27. Kanis E, van den Belt H, Groen AF, Schakel J, de Greef KH (2004) Breeding for improved welfare in pigs: a conceptual framework and its use in practice. *Anim Sci* 78:315–329
28. Clark JAM, Potter M, Harding E (2006) The welfare implications of animal breeding and breeding technologies in commercial agriculture. *Livest Sci* 103:270–281
29. Mark T, Sandøe P (2010) Genomic dairy cattle breeding: risks and opportunities for cow welfare. *Anim Welfare* 19(1):113–121
30. Meuwissen THE (1997) Maximising the response of selection with a predefined rate of inbreeding. *J Anim Sci* 75: 934–940

### Books and Reviews

- FAO (2010) Breeding strategies for sustainable management of animal genetic resources: FAO animal production and health guidelines, vol 3. FAO, Rome, 160 pp

## Animal Breeding, Foundations of

GUILHERME J. M. ROSA

Departments of Animal Sciences, Biostatistics & Medical Informatics, Dairy Science, University of Wisconsin, Madison, WI, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Principles of Selection  
Mixed Model Methodology  
Marker-Assisted Selection  
Future Directions  
Bibliography

### Glossary

**Bayesian inference** Statistical inference approach based on the combination of prior information and evidence (i.e., observations) for estimation or hypothesis testing. In Bayesian analysis the prior information is updated with the experimental data to generate the posterior distribution of unknowns, such as model parameters. The name “Bayesian” comes from the use of the Bayes’ theorem in the updating process.

**Breeding value** A measure of the genetic merit of an individual for breeding purposes.

**Genetic correlation** The correlation between traits that is caused by genetic as opposed to environmental factors. Genetic correlations can be caused by pleiotropy (genes that affect multiple traits simultaneously) or by linkage disequilibrium between genes affecting the different traits.

**Genomic selection** Genomic selection is a form of marker-assisted selection in which genetic markers covering the whole genome are used such that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker.

**Heritability (narrow sense)** The fraction of the phenotypic variance that is due to additive genetic effects.

**Infinitesimal genetic model** A genetic model that assumes that a trait is influenced by a very large

(effectively infinite) number of loci, each with infinitesimal effect.

**Linkage disequilibrium** Nonrandom association of alleles at two or more loci, leading to combinations of alleles (haplotypes) that are more or less frequent in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies.

**Mixed models** A mixed-effects model (or simply mixed model) is a statistical model containing both fixed and random effects. Such models are useful in a wide variety of disciplines in the physical, biological, and social sciences, especially for the analysis of data with repeated measurements on each statistical unit or with measurements taken on clusters of related statistical units.

**Population genetics** The study of allele frequency distribution and change under the influence of the four main evolutionary processes: selection, genetic drift, mutation, and migration.

**Quantitative genetics** The study of complex traits (e.g., production and reproductive traits, disease resistance) and their underlying genetic mechanisms. It is effectively an extension of simple Mendelian inheritance in that the combined effect of the many underlying genes results in a continuous distribution of phenotypic values or of some underlying scale or liability thereof.

### Definition of the Subject

The term *Animal Breeding* refers to the human-guided genetic improvement of phenotypic traits in domestic animals such as livestock and companion species [1]. Animal breeding is based on principles of *Quantitative Genetics* [2–4] and aims to increase the frequency of favorable alleles and allelic combinations in the population, which is achieved through selection of superior individuals and specific mating systems strategies. Selection methods and mating strategies are developed by combining principles of quantitative and population genetics with sophisticated statistical methods and computational algorithms for integrating phenotypic, pedigree, and genomic information, along with the utilization of reproductive technologies that allow for larger progeny cohorts from superior animals as well as shorter generation intervals.



Through selection and mating of superior animals the frequency of favorable alleles is increased, so the overall additive genetic merit of a population is increased through successive generations [5]. Selection can be regarded as the most important tool for the improvement of lines or breeds within a specific species in terms of additive genetic effects. Such lines or breeds can be then intermated such that nonadditive genetic effects such as dominance and epistasis can be exploited through specific inter- and intralocus allelic combinations [1–4].

The theoretical foundations of population and quantitative genetics can be traced back to the work of R. A. Fisher, J. B. S. Haldane, and S. Wright. The rational animal breeding has its origins in the work of J. L. Lush, who made substantial contributions to animal genetics and biometrics research and is generally referred to as the father of modern scientific animal breeding [1].

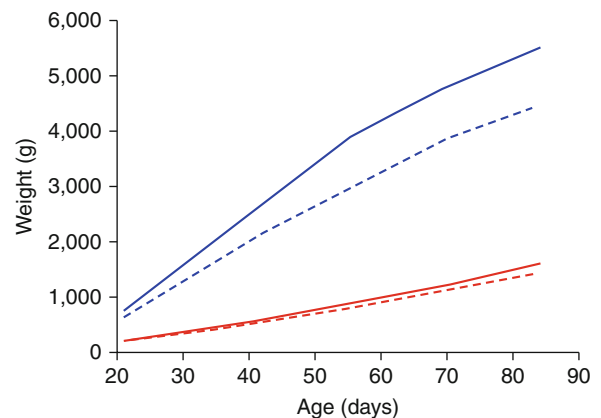
More recent theoretical developments in population and quantitative genetics have been fostered by researchers such as C. C. Cockerham, C. W. Cotterman, J. F. Crow, W. J. Ewens, W. G. Hill, M. Kimura, G. Malécot, T. Nagylaki, and B. S. Weir, among others. A landmark in the area of animal breeding and genetics is the development of mixed model methodology, first proposed by C. R. Henderson, which has been used extensively in many applications in the field, ranging from breeding value prediction under the infinitesimal assumption to gene mapping and segregation analysis. Most recently, Bayesian methods, Monte Carlo, and resampling techniques have been employed to fit and evaluate complex models in different contexts, including nonlinear systems, generalized models, survival analysis, and situations in which the number of parameters or covariates surpasses the number of observations, such as in association analysis and whole-genome marker-assisted selection using high density panels of single nucleotide polymorphism (SNP) markers.

## Introduction

Since domestication, artificial selection has greatly changed the shape, size, and production and reproduction performance of livestock and companion animal species. For example, there is an incredible diversity of

canine breeds – and between dogs and their wolf ancestors – from differences in overall appearance to behavior and their ability to perform specific tasks. Although to a lesser degree, the same can be observed in many other companion animal species, such as cats and horses. With livestock species, tremendous genetic changes have been accomplished as well, markedly in the last 50 years or so. For example, Fig. 1 depicts the average growth curves of broilers from selected and control populations. These results refer to a population of birds selected for over 40 years for increased growth rate and another population kept without artificial selection, with both groups derived from the same base population, starting in 1957 [6]. In the experiment presented in Fig. 1, the two groups of birds were fed diets typical of 1957 and 2001, such that the interaction between genetics and feed, as well as the genetic contribution to the phenotypic differences observed, could be assessed. It is seen that the 2001 genetics group presented an average body weight of about 4 kg at 56 days of age, while its 1957 counterpart weighed only 800 g or so. Moreover, it is shown that 85–90% of this fivefold improvement is accounted for by genetics with the remaining 10–15% to nutrition.

Similar levels of genetic improvement can also be observed in many other species, such as swine, beef and



**Animal Breeding, Foundations of. Figure 1**

Average growth curves of commercial broilers. *Blue and red lines* represent birds with “2001” and “1957” genetics, respectively. *Solid and dashed lines* represent birds fed diets typical of 2001 or 1957, respectively (Adapted from [6])

dairy cattle, and some species of fish. For example, as illustrated in Fig. 2, the average breeding value for milk yield in the US Holstein, and Red and White populations has increased over 3,500 kg in the last 50 years.

Such genetic improvements have been accomplished mostly through the selection and breeding of superior animals, which can be chosen using specific statistical methods such as those discussed on the following sections. In this chapter, the discussion will focus on methods developed for normally distributed (Gaussian) traits, under the infinitesimal assumption, i.e., that traits are affected by a large (virtually infinite) number of genes of small effects [2–4], although this assumption is somewhat alleviated in marker-assisted selection, which is discussed later.

## Principles of Selection

### Basic Genetic Model for Quantitative Traits

The basic genetic model can be expressed as [2, 3, 7]:

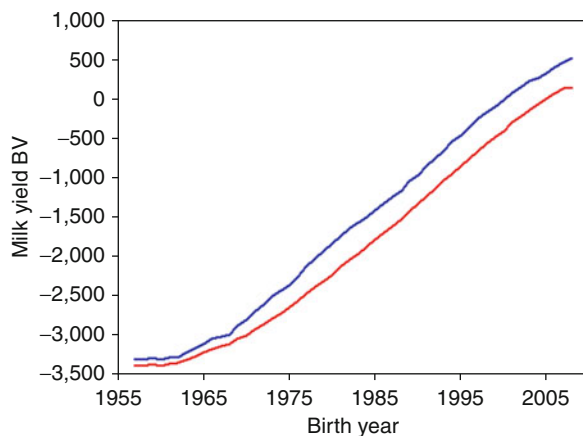
$$y_i = \mu + g_i + e_i \quad (1)$$

where  $y_i$  is the phenotypic value of animal  $i$  (i.e., the animal's performance for a specific trait);  $\mu$  is the

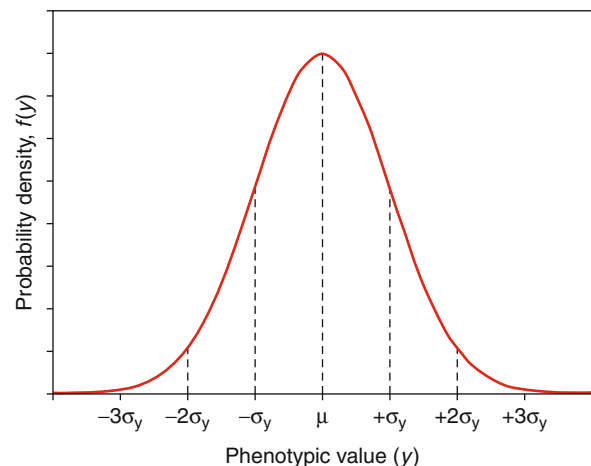
population mean (average performance of the animals);  $g_i$  is the genotypic value of the animal, expressed as a deviation from the mean; and  $e_i$  is a term representing environmental factors affecting the animal's performance, also expressed as a deviation from the mean. Hence, it is assumed that  $E[g_i] = 0$  and  $E[e_i] = 0$ , such that  $E[y_i] = \mu$ , where  $E[\cdot]$  represents the expectation function. Moreover, the variance of  $y_i$  is given by  $\text{Var}[y_i] = \sigma_y^2 = \sigma_g^2 + \sigma_e^2$ , where  $\sigma_g^2 = \text{Var}[g_i]$  and  $\sigma_e^2 = \text{Var}[e_i]$  are the genetic and environmental variances, respectively. Normally distributed traits, i.e., phenotypic traits with a bell-shaped distribution, are generally represented as  $y_i \sim N(\mu, \sigma_y^2)$ . Such distribution has a probability density function that can be described as [2, 4]:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left\{-\frac{1}{2\sigma_y^2}(y_i - \mu)^2\right\},$$

for  $-\infty < y_i < \infty$ ,  $-\infty < \mu < \infty$ , and  $\sigma_y^2 > 0$ , which can be represented as in Fig. 3. To simplify the notation used throughout the text, it is noted that either random variables or their realizations will be represented with lower case letters. However, the context should make it clear to the reader when a letter represents one or the other.



**Animal Breeding, Foundations of. Figure 2** Genetic trend for milk yield in the US Holstein and Red and White populations. Males and females average breeding values are in blue and red, respectively; genetic base refers to cows born in year 2005 (Source: AIPL – USDA; <http://www.aipl.arsusda.gov/>)



**Animal Breeding, Foundations of. Figure 3** Probability density function of a normally distributed trait with mean  $\mu = E[y_i]$  and variance  $\sigma_y^2 = \text{Var}[y_i]$ , i.e.,  $y_i \sim N(\mu, \sigma_y^2)$

The genetic component  $g_i$  of Model (1) can be partitioned into additive ( $a_i$ ) and nonadditive ( $c_i$ ) genetic effects, i.e.,  $g_i = a_i + c_i$ , where  $a_i$  is also called “breeding value,” and  $c_i$  refers to the “gene combination value,” which encompasses interaction effects between alleles within each locus (i.e., dominance effects) or between alleles in different loci (i.e., epistatic effects).

Hence, Model (1) can be expressed as:

$$y_i = \mu + a_i + c_i + e_i \quad (2)$$

where  $a_i \sim N(0, \sigma_a^2)$ ,  $c_i \sim N(0, \sigma_c^2)$ , and  $e_i \sim N(0, \sigma_e^2)$ , with all these terms assumed independent from each other. The phenotypic variance can be then expressed as  $\text{Var}[y_i] = \sigma_y^2 = \sigma_a^2 + \sigma_c^2 + \sigma_e^2$ , from which two important definitions are derived. The first one is called broad sense heritability, given by  $H^2 = \sigma_g^2 / \sigma_y^2$ , where  $\sigma_g^2 = \sigma_a^2 + \sigma_c^2$ , which gives the proportion of the phenotypic variance that is due to genetic effects. The second, called narrow sense heritability, refers to the specific contribution of additive genetic effects to the phenotypic variance, i.e.,  $h^2 = \sigma_a^2 / \sigma_y^2$ . These two quantities, particularly narrow sense heritability, will be further discussed and used in the next sections.

The breeding value of an individual ( $a_i$ ) is equal to the sum of additive effects of individual alleles within and across loci, and it is sometimes called “additive genetic deviation” or “additive genetic effect.” Because individual alleles, and therefore independent allele effects, are passed from parent to offspring, the breeding value of an individual is important for predicting its progeny’s performance and so it is central to selection of superior animals [1, 3]. The gene combination value ( $c_i$ ) is the difference between the genetic merit ( $g_i$ ) of an animal and its breeding value, i.e.,  $c_i = g_i - a_i$ , so it is often called “nonadditive genetic deviation.” Because the component  $c_i$  involves interactions between alleles (both within and between loci), and only a single allele (as opposed to a pair of alleles) in each locus is transmitted from parents to offspring, nonadditive effects are not transmitted in a predictable manner. Hence, while average breeding value in a population can be changed through selection of superior animals, the gene combination value should be explored through specific mating systems. Here, the discussion will focus on selection approaches and the genetic improvement of a population in terms of

additive genetic effects only. For a discussion on mating systems, such as inbreeding and outbreeding strategies, see for example, [1, 7, 8]. Additional discussion on inbreeding depression and heterosis (or hybrid vigor) can be found in [3, 4].

As discussed previously, the breeding value of an individual is equal to the sum of its independent allele effects. Because a parent passes a random sample of half of its alleles to its progeny, an animal’s breeding value is twice what is often called “transmitting ability” or “expected progeny difference” [1, 5]. The expected breeding value of an offspring ( $a_o$ ) is then equal to the average of its parents’ breeding values (the same as the sum of its parents’ transmitting abilities), i.e.,  $E[a_o | a_s, a_d] = \frac{a_s + a_d}{2}$ , where  $a_s$  and  $a_d$  represent the (realized) breeding values of the offspring’s sire and dam, respectively. However, there will be variability in terms of breeding values within a full-sib family because of the random sampling of parents’ alleles that each offspring receives, the so-called Mendelian sampling [4].

The breeding value of an individual can be expressed as a function of its parents’ breeding values as  $a_o = 0.5a_s + 0.5a_d + \delta$ , where  $\delta$  refers to the Mendelian sampling component. It is interesting to notice that the variance of breeding values in a specific generation is equal to  $\text{Var}[a_o] = 0.25\text{Var}[a_s] + 0.25\text{Var}[a_d] + \text{Var}[\delta]$ . Assuming the same additive genetic variance across generations and for both sexes (i.e.,  $\text{Var}[a_o] = \text{Var}[a_s] = \text{Var}[a_d] = \sigma_a^2$ ), it is shown that the Mendelian sampling variance is equal to half the additive genetic variance, i.e.,  $\text{Var}[\delta] = \sigma_a^2/2$ .

## Phenotypic Selection

The most traditional approach of genetic improvement of livestock (and more generally any domestic animal or plant species) is based on selection of animals with the best performance, or “phenotypic selection” [1–4]. Accordingly, given a group of animals supposedly reared in similar environmental conditions, only those with the highest performance are allowed to breed to produce the next generation. As discussed previously (Model 2), the performance of each animal is a combination of its breeding value and all other nonadditive genetic effects and environmental factors, such that a superior performance does not always

represent superior breeding value. Nonetheless, whenever  $\sigma_a^2 > 0$ , there will be a positive correlation between performance and breeding value, and the effectiveness of phenotypic selection (i.e., selection response) will increase with such correlation.

To illustrate this concept consider Fig. 4, in which a scatter plot of breeding values and phenotypes (centered on zero, i.e.,  $y_i - \mu$ ) for a few fictitious animals is presented. As indicated before, in this chapter the discussion will be focused on selection approaches and the genetic improvement of a population in terms of additive genetic effects only, such that Model (2) can be conveniently reexpressed as:

$$y_i = \mu + a_i + \varepsilon_i \quad (3)$$

where  $\varepsilon_i = c_i + e_i$  represents all nonadditive genetic and environmental effects affecting the phenotypic value  $y_i$ , assumed  $\varepsilon_i \sim N(0, \sigma_e^2)$ .

Assuming that each effect in Model (3) is independent from each other, the covariance between phenotype and breeding value is given by:

$$\text{Cov}[y_i, a_i] = \text{Cov}[\mu + a_i + \varepsilon_i, a_i] = \text{Var}[a_i] = \sigma_a^2,$$

such that the correlation between phenotype and breeding value is:

$$r_{y_i, a_i} = \frac{\text{Cov}[y_i, a_i]}{\sqrt{\text{Var}[y_i]\text{Var}[a_i]}} = \frac{\sigma_a^2}{\sigma_y \sigma_a} = \frac{\sigma_a}{\sigma_y} = \sqrt{h^2},$$

i.e., the square root of the (narrow sense) heritability.

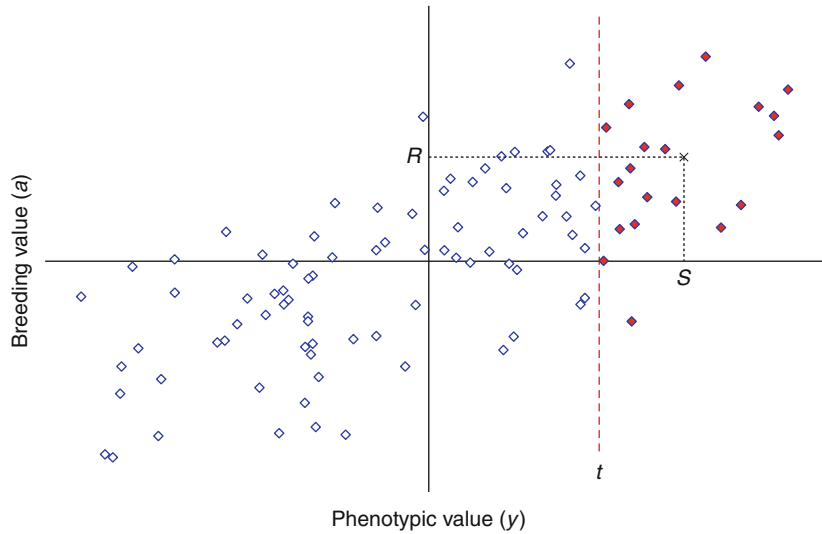
As the breeding values of animals are unknown in practice, what phenotypic selection does is to predict (or estimate) the animals' breeding values based on their own performance. The prediction is based on the regression of breeding values on phenotypes, and the regression coefficient (slope) is given by:

$$b_{a_i \bullet y_i} = \frac{\text{Cov}[y_i, a_i]}{\text{Var}[y_i]} = \frac{\sigma_a^2}{\sigma_y^2} = h^2.$$

This means that an animal's estimated breeding value (EBV) based solely on its performance (and with a single measurement only) can be expressed as:

$$\hat{a}_i = h^2(y_i - \mu).$$

The correlation between such EBV (which is a linear transformation of  $y_i$ ) and the true breeding



**Animal Breeding, Foundations of. Figure 4**

Scatter plot of breeding values versus phenotypic values. Each dot represents a specific animal and those colored in red are selected animals with performance (i.e., phenotypic value) above a specified threshold ( $t$ ).  $S$  and  $R$  represent the average phenotypic and breeding values of the selected (*top*) animals, respectively

value ( $a_i$ ) is  $r_{\hat{a}_i, a_i} = \frac{\text{Cov}[\hat{a}_i, a_i]}{\sqrt{\text{Var}[\hat{a}_i]\text{Var}[a_i]}} = \frac{h^2 \sigma_a^2}{\sqrt{h^4 \sigma_y^2 \sigma_a^2}} = h$ , which is generally referred to as “prediction accuracy” in the animal breeding literature [5]. The square of the accuracy in this case is equal to the heritability of the trait and is often called “prediction reliability.” The prediction accuracy (and consequently the reliability) can be increased by using additional sources of information on an animal (such as repeated measurements of the trait or performance of progeny and other relatives) when estimating its breeding value. An example is with selection indexes and mixed model methodology, which will be discussed later in this chapter.

As indicated in Fig. 4, the selected animals (i.e., the best performing animals) will have an average phenotypic value equal to  $S$  and an average breeding value equal to  $R$ . The expected average breeding value (and also the expected phenotypic performance) of the progeny of the selected animals is also  $R$ , as illustrated in Fig. 5, and the ratio  $R/S$  is equal to the heritability ( $h^2$ ) of the trait under selection. The genetic progress after one generation of selection is then given by:

$$R = h^2 S,$$

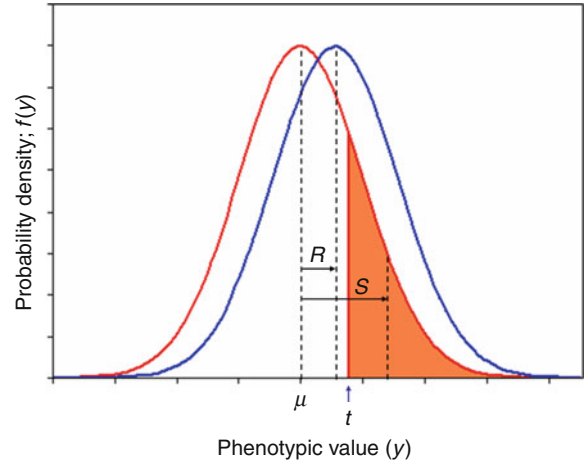
where  $R = \mu_p - \mu$  and  $S = \mu_s - \mu$ , with  $\mu_p$ ,  $\mu_s$ , and  $\mu$  representing the average phenotypic performance of the progeny (generation 1), of the selected animals, and of the selection candidate (generation 0) populations, respectively.

The selection differential ( $S$ ) can also be expressed as  $S = i\sigma_y$ , where  $i = \frac{\mu_s - \mu}{\sigma_y}$  is called “selection intensity,” and represents the selection differential in terms of phenotypic standard deviations. In addition, as  $R$  represents the genetic progress expected in a single generation of selection, the genetic improvement per unit of time is then given by  $R^* = R/L$ , where  $L$  is the generation interval. Hence, the expected genetic progress when phenotypic selection on a single trait is employed is [1, 3]:

$$R^* = \frac{h^2 i \sigma_y}{L},$$

which, given that  $\sigma_y = \sigma_a/h$ , can be expressed also as:

$$R^* = \frac{hi\sigma_a}{L}.$$



**Animal Breeding, Foundations of. Figure 5**

Probability density of the distribution of phenotypic values in the candidates-for-selection (red) and the progeny (blue) populations. The candidates-for-selection group represents the parental population (or generation 0), from which the top performing animals (above the threshold  $t$ ) are selected and mated to produce the next generation, or progeny (generation 1). The difference between the phenotypic average of the selected animals and that of the generation 0 is called selection differential (represented by  $S$ ), and the difference between the phenotypic mean of the progeny and that of the generation 0 is called genetic progress, or genetic response (represented by  $R$ )

This equation is a special form of the so-called “breeder’s equation” (or “key equation”), for the case of phenotypic selection. In its general form, the breeder’s equation is expressed as [5]:

$$R^* = \frac{\text{Accuracy} \times \text{Intensity} \times \text{Variation}}{\text{Generation interval}},$$

meaning that the genetic progress per unit of time is proportional to the accuracy of breeding values prediction, to the selection intensity, and to the genetic variation, and inversely proportional to the generation interval.

Hence, to increase the genetic progress in a population (e.g., breed or line) through selection, animal breeders (and similarly plant breeders) work to improve the four components of the equation above. As the genetic variability is a natural characteristic of a population and cannot be easily changed, genetic



progress is generally incremented by improving prediction accuracy (e.g., by using specific statistical techniques to combine different sources of information regarding the animals' genetic merit), by increasing the selection intensity, and by shortening the generation interval, which can be accomplished using molecular genetics techniques (e.g., the use of marker-assisted selection) and biotechnology approaches (e.g., artificial insemination).

It is important to mention that the breeder's equation discussed here can be extended for more complex scenarios, such as when males and females contribute differently for some components of the equation [5]. For example, prediction accuracies and selection intensity are generally higher for males if artificial insemination is used. Another important issue to mention here is that selection not only shifts the mean of the breeding values in a population but also changes the genetic variance (and heritability). A primary cause of the change in genetic variance is due to the fact that selected parents represent one tail of the phenotypic distribution, therefore their phenotypic variance is smaller than that of the whole candidates-for-selection population. This leads to a reduction in both the phenotypic and additive genetic variances in the progeny population, which is known as the "Bulmer effect" [2]. In addition, as selection modifies allele frequencies toward the fixation of favorable alleles, selection in one direction over many generations is also expected to reduce the genetic variation. Additional discussion on effects of selection on variance and other short-term and long-term consequences of artificial selection can be found, for example, in [2, 3].

In the remainder of this chapter specific statistical techniques (such as the selection index, BLUP, and genomic selection) for the improvement of accuracy, intensity, and generation interval, and consequently the increase of genetic progress from artificial selection, will be discussed.

### Correlated Response and Indirect Selection

If two traits  $x$  and  $y$  are genetically correlated, direct selection on one of the traits (say  $y$ ) will also cause a genetic change in the other trait (trait  $x$ ), which is called "correlated response" [3]. Correlated response to

selection ( $R_{x\bullet y}$ ), that is, genetic change in trait  $x$  as a consequence of direct selection on trait  $y$ , can be predicted by:

$$R_{x\bullet y} = b_{x\bullet y} R_y,$$

where  $R_y$  is the genetic progress of trait  $y$  through direct selection on itself, and  $b_{x\bullet y}$  is the genetic regression coefficient, given by:

$$b_{x\bullet y} = \frac{\text{Cov}(a_x, a_y)}{\sigma_{a_y}^2},$$

where  $\text{Cov}(a_x, a_y)$  is the genetic covariance between traits  $x$  and  $y$ .

The genetic correlation between two traits  $x$  and  $y$  is given by:

$$\rho_{a_x, a_y} = \frac{\text{Cov}(a_x, a_y)}{\sigma_{a_x} \sigma_{a_y}},$$

such that  $\text{Cov}(a_x, a_y) = \rho_{a_x, a_y} \sigma_{a_x} \sigma_{a_y}$  and the genetic regression can be expressed as:

$$b_{x\bullet y} = \frac{\rho_{a_x, a_y} \sigma_{a_x} \sigma_{a_y}}{\sigma_{a_y}^2} = \rho_{a_x, a_y} \frac{\sigma_{a_x}}{\sigma_{a_y}}.$$

Using this term, and recalling the selection response formula discussed before, given by  $R_y = h_y i_y \sigma_{a_y}$ , the correlated response can then be expressed as:

$$R_{x\bullet y} = \rho_{a_x, a_y} \frac{\sigma_{a_x}}{\sigma_{a_y}} h_y i_y \sigma_{a_y} = \rho_{a_x, a_y} \sigma_{a_x} h_y i_y,$$

or, given that  $\sigma_{a_x} = h_x \sigma_{y_x}$ , it can be finally written as:

$$R_{x\bullet y} = \rho_{a_x, a_y} h_x h_y i_y \sigma_{y_x}.$$

Such an equation can be used either to monitor potential genetic changes in correlated traits when performing direct selection on a specific trait of economic importance or, alternatively, to explore indirect selection strategies using indicator traits [5]. The latter use may be of interest when a trait of economic importance (e.g., trait  $x$ ) is difficult or expensive to measure, or it is expressed later in an animal's life, so it may be advantageous to select on a correlated trait (e.g., trait  $y$ ), which would be the indicator trait. To assess the effectiveness of indirect selection relative to direct selection, one may look at

the ratio of expected genetic progress per unit of time in each scenario, i.e.,

$$\begin{aligned} \frac{R_{x \bullet y}}{R_x} &= \frac{\rho_{a_x, a_y} \sigma_{a_x} h_y i_y / L_y}{h_x i_x \sigma_{a_x} / L_x} = \frac{\rho_{a_x, a_y} h_y i_y L_x}{h_x i_x L_y} \\ &= \rho_{a_x, a_y} \frac{h_y i_y L_x}{h_x i_x L_y}. \end{aligned}$$

So, it can be seen that this ratio can be higher than 1 (meaning that the indirect selection is more effective than the direct selection) depending on the genetic correlation between the economic and the indicator traits, the ratios of their heritabilities, and their potential selection intensities and generation intervals.

### Selection Index

In section “[Phenotypic Selection](#),” selection based on a single measurement on each animal was discussed. However, it is not always possible to observe the phenotype for all animals, such as traits that are expressed in only one sex or that require the sacrifice of animals to be measured, etc. In addition, even when it is possible to measure the phenotypic trait in each animal, information from relatives can be used to obtain earlier or more reliable predictions of breeding values. In this section, the prediction of breeding values using different sources of information (e.g., multiple measurements of the trait in each animal and progeny performance) will be discussed and a methodology (the selection index) that combines multiple sources of information into a single prediction for each animal will be presented.

When multiple measurements of the same trait are recorded (e.g., milk yield in multiple lactations), breeding values can be predicted using the average of observations ( $\bar{y}_i$ ) from each animal as  $\hat{a}_i = b_{a_i \bullet \bar{y}_i}(\bar{y}_i - \mu)$ . However, to derive the genetic regression of breeding value on average phenotypic value, Model (3) must be expanded to include an additional term, which is discussed next.

It can be shown empirically that the covariance (or resemblance) between repeated measurements on the same animal is larger than  $\sigma_a^2$ , which is what would be expected under the assumptions of Model (3). This additional source of covariance between records for the same animal refers to environmental factors that

affect all records similarly, the so-called “permanent environmental effects” [1, 4]. Under these circumstances, the Model (3) can be extended to:

$$y_{ij} = \mu + a_i + p_i + \varepsilon_{ij} \quad (4)$$

where  $y_{ij}$  represents the observation  $j$  ( $j = 1, \dots, n_i$ ) on animal  $i$ , with  $n_i$  being the total number of records on animal  $i$ ;  $\mu$  and  $a_i \sim N(0, \sigma_a^2)$  are as defined previously;  $p_i$  refers to the permanent environmental effects affecting records on animal  $i$ , assumed  $p_i \sim N(0, \sigma_p^2)$ ; and  $\varepsilon_{ij} \sim N(0, \sigma_e^2)$  represent residual effects (nonadditive genetic and temporary environmental effects) associated with observation  $y_{ij}$ . In addition, it is assumed that all random terms in Model (4) are independent from each other, i.e.,  $\text{Cov}(a_i, p_i) = \text{Cov}(a_i, \varepsilon_{ij}) = \text{Cov}(p_i, \varepsilon_{ij}) = \text{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = 0$  for any  $i, j$ , and  $j' (j \neq j')$ .

Under these settings, the average phenotypic value of an animal is given by  $\bar{y}_i = \mu + a_i + p_i + \bar{\varepsilon}_i$ , where

$\bar{\varepsilon}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \varepsilon_{ij}$ , such that its variance is given by

$\text{Var}[\bar{y}_i] = \sigma_a^2 + \sigma_p^2 + \sigma_e^2/n_i$ , and the covariance between  $a_i$  and  $\bar{y}_i$  is  $\text{Cov}(a_i, \bar{y}_i) = \sigma_a^2$ . In this case, the regression of breeding values on phenotypic means is given by:

$$b_{a_i \bullet \bar{y}_i} = \frac{\text{Cov}[a_i, \bar{y}_i]}{\text{Var}[\bar{y}_i]} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_p^2 + \sigma_e^2/n_i}.$$

An important definition related to repeated measurements refers to repeatability ( $r$ ), which is given by the intraclass correlation, i.e., the ratio of the within-individual (or between repeated measurements) to the phenotypic variances [1, 4]:

$$r = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_y^2} = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_a^2 + \sigma_p^2 + \sigma_e^2},$$

and measures the correlation between records on the same animal.

Noting that  $r = 1 - \frac{\sigma_e^2}{\sigma_a^2 + \sigma_p^2 + \sigma_e^2}$ , the variance of the average phenotypic value of an animal can be expressed as a function of the repeatability as  $\text{Var}[\bar{y}_i] = [r + (1 - r)/n_i] \sigma_y^2$ , such that the genetic regression becomes:

$$b_{a_i \bullet \bar{y}_i} = \frac{\sigma_a^2}{[r + (1 - r)/n_i] \sigma_y^2} = \frac{n_i h^2}{1 + (n_i - 1)/r}.$$

The prediction accuracy in this case, i.e., the correlation between an animal's estimated breeding value

using repeated records and its true breeding value is given by:

$$\begin{aligned} r_{\hat{a}_i, a_i} &= r_{\bar{y}_i, a_i} = \frac{\text{Cov}(\bar{y}_i, a_i)}{\sqrt{\text{Var}(\bar{y}_i)\text{Var}(a_i)}} \\ &= \frac{\sigma_a^2}{\sqrt{r + (1-r)/n_i}\sigma_y\sigma_a} \\ &= h\sqrt{\frac{n_i}{1 + (n_i - 1)r}} = \sqrt{b_{a_i \bullet \bar{y}_i}}. \end{aligned}$$

Hence, it can be seen that compared with single record phenotypic selection, there is a gain in accuracy when predictions are based on repeated records and that the gain will depend on the values of  $r$  and  $n_i$ ; higher gain in accuracy is obtained when  $r$  is low and when  $n_i$  is high.

Another alternative to predict breeding values is to use progeny performance, which is often employed for predicting breeding values of males for traits where records can be obtained only on females, such as milk yield. For example, let  $\bar{y}_i$  be the average of single records on  $n_i$  progeny of sire  $i$ , and assume that the sire was mated to a random sample of females not related to him. In this case, each progeny record can be expressed as:

$$y_{ij} = \mu + \frac{1}{2}a_i + \frac{1}{2}d_{ij} + \delta_{ij} + \varepsilon_{ij},$$

where  $a_i$  is the breeding value of a specific sire  $i$ ;  $d_{ij}$  is the breeding value of dam  $j$  ( $j = 1, \dots, n_i$ ) mated with sire  $i$ ; and  $\delta_{ij}$  and  $\varepsilon_{ij}$  refer to the Mendelian sampling and residual (environmental) components associated with the observation  $y_{ij}$ . Using this notation, the following model can be used to describe the progeny average of sire  $i$ :

$$\bar{y}_i = \mu + \frac{1}{2}a_i + \frac{1}{2}\bar{d}_i + \bar{\delta}_i + \bar{\varepsilon}_i \quad (5)$$

where  $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ ,  $\bar{d}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} d_{ij}$ ,  $\bar{\delta}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{ij}$ , and  $\bar{\varepsilon}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \varepsilon_{ij}$ . Given that  $E[\bar{d}_i] = 0$  and  $E[\bar{\delta}_i] = 0$ , the breeding value of sire  $i$  can be then predicted by  $\hat{a}_i = b_{a_i \bullet \bar{y}_i}(\bar{y}_i - \mu)$ , where  $b_{a_i \bullet \bar{y}_i} = \text{Cov}[a_i, \bar{y}_i]/\text{Var}[\bar{y}_i]$ . It is shown that:

$$\text{Cov}(a_i, \bar{y}_i) = \text{Cov}(a_i, a_i/2) = \sigma_a^2/2$$

and

$$\begin{aligned} \text{Var}[\bar{y}_i] &= \text{Var}\left[\frac{1}{2}a_i + \frac{1}{2}\bar{d}_i + \bar{\delta}_i + \bar{\varepsilon}_i\right] \\ &= \frac{1}{4}\sigma_a^2 + \frac{1}{4}\frac{\sigma_a^2}{n_i} + \frac{\sigma_a^2}{2n_i} + \frac{\sigma_\varepsilon^2}{n_i} \\ &= \frac{(n_i + 3)\sigma_a^2 + 4\sigma_\varepsilon^2}{4n_i} \\ &= \frac{(n_i + 3)h^2 + 4(1 - h^2)}{4n_i}\sigma_y^2 \\ &= \left[k + \frac{1 - k}{n_i}\right]\sigma_y^2, \end{aligned}$$

where  $k = h^2/4$  is the intraclass correlation between half-sibs, such that the genetic regression coefficient is given by:

$$\begin{aligned} b_{a_i \bullet \bar{y}_i} &= \frac{\sigma_a^2/2}{[k + (1 - k)/n_i]\sigma_y^2} \\ &= \frac{h^2\sigma_y^2/2}{[h^2/4 + (1 - h^2/4)/n_i]\sigma_y^2} \\ &= \frac{2n_i h^2}{4 + (n_i - 1)h^2}, \end{aligned}$$

and the prediction accuracy by:

$$\begin{aligned} r_{a_i, \bar{y}_i} &= \frac{\text{Cov}[a_i, \bar{y}_i]}{\sqrt{\text{Var}[a_i]\text{Var}[\bar{y}_i]}} = \frac{h^2\sigma_y^2/2}{\sqrt{h^2\sigma_y^2[k + (1 - k)/n_i]\sigma_y^2}} \\ &= \sqrt{\frac{n_i h^2/4}{1 + (n_i - 1)k}} \\ &= \sqrt{\frac{n_i h^2}{4 + (n_i - 1)h^2}} = \sqrt{b_{a_i \bullet \bar{y}_i}/2}, \end{aligned}$$

which approaches unity (one) as the number of progeny records increases.

Up to this point, it has been discussed how breeding values can be predicted using different sources of information, such as an animal's own performance (either a single record or multiple measurements) or progeny performance. Other sources of information that could also be used are the performance of parents, sibling, or other kinds of relatives. However, what generally happens is that multiple sources of information are available simultaneously, so the question becomes how to best combine all the information in order to improve prediction accuracy. Here, a classical approach will be discussed, the "selection index," and later on in this

chapter a more general and modern alternative, based on mixed model methodology, will be presented.

Consider, for example, that there are three sources of information available on animal  $i$  (represented here as  $y_{i1}$ ,  $y_{i2}$ , and  $y_{i3}$ , and expressed as deviations from their means), so the goal is to predict the animal's breeding value with a linear combination of such information, i.e.,

$$\hat{a}_i = b_{i1}y_{i1} + b_{i2}y_{i2} + b_{i3}y_{i3},$$

such that the prediction accuracy (i.e., correlation between predicted and true breeding value) is maximized.

Maximization of  $r_{\hat{a}_i, a_i}$  is equivalent to the maximization of  $\log(r_{\hat{a}_i, a_i})$ , which is generally easier to accomplish. The log correlation can be expressed as (here, to simplify the notation, the index  $i$  indicating the animal is dropped):

$$\log(r_{\hat{a}, a}) = \log \left[ \frac{\sigma_{\hat{a}, a}}{\sqrt{\sigma_{\hat{a}}^2 \sigma_a^2}} \right] = \log(\sigma_{\hat{a}, a}) - \frac{1}{2} \sigma_{\hat{a}}^2 - \frac{1}{2} \sigma_a^2,$$

where the covariance between  $\hat{a}$  and  $a$ , and the variance of  $\hat{a}$  are given respectively by:

$$\sigma_{\hat{a}, a} = b_1 \sigma_{y_1, a} + b_2 \sigma_{y_2, a} + b_3 \sigma_{y_3, a}$$

and

$$\sigma_{\hat{a}}^2 = b_1^2 \sigma_{y_1}^2 + 2b_1 b_2 \sigma_{y_1, y_2} + 2b_1 b_3 \sigma_{y_1, y_3} + b_2^2 \sigma_{y_2}^2 + 2b_2 b_3 \sigma_{y_2, y_3} + b_3^2 \sigma_{y_3}^2.$$

Substituting these expressions into  $\log(r_{\hat{a}, a})$ , taking the partial derivatives of  $\log(r_{\hat{a}, a})$  with respect to each of the regression coefficients  $b_j$  ( $j = 1, 2, 3$ ), and setting them to zero, gives the following set of equations:

$$\begin{cases} \frac{\partial \log(r_{\hat{a}, a})}{\partial b_1} = \frac{\sigma_{y_1, a}}{\sigma_{\hat{a}, a}} - \frac{b_1 \sigma_{y_1}^2 + b_2 \sigma_{y_1, y_2} + b_3 \sigma_{y_1, y_3}}{\sigma_{\hat{a}}^2} \\ \frac{\partial \log(r_{\hat{a}, a})}{\partial b_2} = \frac{\sigma_{y_2, a}}{\sigma_{\hat{a}, a}} - \frac{b_1 \sigma_{y_1, y_2} + b_2 \sigma_{y_2}^2 + b_3 \sigma_{y_2, y_3}}{\sigma_{\hat{a}}^2} \\ \frac{\partial \log(r_{\hat{a}, a})}{\partial b_3} = \frac{\sigma_{y_3, a}}{\sigma_{\hat{a}, a}} - \frac{b_1 \sigma_{y_1, y_3} + b_2 \sigma_{y_2, y_3} + b_3 \sigma_{y_3}^2}{\sigma_{\hat{a}}^2} \end{cases}$$

which can be rearranged as:

$$\begin{cases} b_1 \sigma_{y_1}^2 + b_2 \sigma_{y_1, y_2} + b_3 \sigma_{y_1, y_3} = k \sigma_{y_1, a} \\ b_1 \sigma_{y_1, y_2} + b_2 \sigma_{y_2}^2 + b_3 \sigma_{y_2, y_3} = k \sigma_{y_2, a} \\ b_1 \sigma_{y_1, y_3} + b_2 \sigma_{y_2, y_3} + b_3 \sigma_{y_3}^2 = k \sigma_{y_3, a} \end{cases}$$

where  $k = \sigma_{\hat{a}}^2 / \sigma_{\hat{a}, a}$ .

Extending the system for any number  $m$  of components (i.e., sources of information), these equations can be expressed in matrix notation as:

$$\mathbf{P}\mathbf{b} = \mathbf{k}\mathbf{c},$$

$$\text{where } \mathbf{P} = \begin{bmatrix} \sigma_{y_1}^2 & \sigma_{y_1, y_2} & \cdots & \sigma_{y_1, y_m} \\ \sigma_{y_1, y_2} & \sigma_{y_2}^2 & \cdots & \sigma_{y_2, y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_1, y_m} & \sigma_{y_2, y_m} & \cdots & \sigma_{y_m}^2 \end{bmatrix} \text{ is the covari-}$$

ance matrix of the vector  $\mathbf{y} = [y_1, y_2, \dots, y_m]'$ ,  $\mathbf{b} = [b_1, b_2, \dots, b_m]'$  is the vector of regression coefficients (weights) of each source of information, and  $\mathbf{c} = [\sigma_{y_1, a}, \sigma_{y_2, a}, \dots, \sigma_{y_m, a}]'$  is the vector of covariances between each piece of information and the breeding value of the animal, such that the weights  $\mathbf{b}$  of the index  $\hat{a} = \mathbf{b}'\mathbf{y}$  are given by  $\mathbf{b} = k\mathbf{P}^{-1}\mathbf{c}$ .

It should be noted that the constant  $k$  does not change the relative size of the regression coefficients  $\mathbf{b}$  or the value of  $r_{\hat{a}, a}$ , so it can be set to 1. In fact, if instead of maximizing  $r_{\hat{a}, a}$ , the average square prediction error  $E[\hat{a} - a]^2$  is minimized, then  $\sigma_{\hat{a}}^2 = \sigma_{\hat{a}, a}$  and the system (usually called selection index equations) becomes:

$$\mathbf{b} = \mathbf{P}^{-1}\mathbf{c}.$$

The correlation between the index and the true breeding value is given by  $r_{\hat{a}, a} = \sigma_{\hat{a}, a} / \sqrt{\sigma_{\hat{a}}^2 \sigma_a^2} = \sqrt{\sigma_{\hat{a}, a} / \sigma_a^2} = \sqrt{\frac{1}{\sigma_a^2} \sum_{j=1}^m b_j \sigma_{y_j, a}}$ .

## Multiple-Trait Selection

Usually more than one trait is considered in a selection program, as multiple traits may be economically important in a production system (e.g., [9]). There are many strategies for multi-trait selection, including the tandem approach (which selects rotationally one trait at a time) and the independent culling levels strategy (which sets minimum performance levels for each of the traits of interest), but they are generally suboptimal.

Here, the selection of a combination of multiple traits evaluated in economic terms will be discussed. Such a combination of traits is generally called "aggregate breeding value" or "breeding objective," and can be expressed as [3]:

$$T = \mathbf{w}'\mathbf{a} = w_1 a_1 + w_2 a_2 + \dots + w_k a_k,$$





The single trait prediction methods discussed in the previous section use only a single source of information or, when multiple sources of information are available, they require them to be split into independent subgroups, i.e., specific groups of relatives such as half-sibs, full-sibs, progeny, etc. However, in practice the data may be extremely complex due to the intricate pedigree structure commonly found in livestock species, e.g., beef and dairy cattle populations. Other drawbacks of the selection index include an inability to account for genetic trend over time and that the phenotypes must be pre-adjusted for environmental effects, which can be done, for example, using the average of contemporary groups of animals. However, contemporary group effects can be inferred only under the unrealistic assumption that they are genetically equal. Hence, a selection index can be reliably applied only to individual animals within same herd and born in same year.

In view of such limitations, linear mixed models (models including both fixed and random effects) and best linear unbiased predictions (BLUP) of breeding values were developed [10–12]. The BLUP methodology uses performance information from all known relatives to estimate breeding values, can be applied to whole herds or large populations using data from many years, and can also accommodate genetic differences between contemporary groups. Presently, mixed models are widely used in many fields of science as a flexible tool for the analysis of data where responses are clustered around some random effects, such that there is a natural dependence between observations in the same cluster. Examples of applications of mixed models in genetics and genomics include gene mapping and association analysis (e.g., [13, 14]), and gene expression assays using microarrays [15, 16] or RT-PCR [17], to name a few.

In some applications of mixed models the central objective is the estimation and hypothesis testing regarding fixed effects (e.g., treatment effects in an experimental study), in which case the random effects (e.g., block effects) are nuisance effects. In animal breeding, however, the main goal is the prediction of realized values of random effects (breeding values of animals), and the fixed effects are generally environmental factors that should be taken into account to adjust the observed phenotypic values. A third application or goal of mixed models is the estimation of

variance components, such as genetic and environmental variances, or functions of them, such as heritability and repeatability.

In this section some basics regarding mixed models are briefly reviewed, with some emphasis toward the prediction of random effects, and subsequently some specific applications of the mixed model methodology in animal breeding and genetics are presented.

A linear mixed-effects model is defined as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (6)$$

where  $\mathbf{y}$  is the vector of responses (observations),  $\boldsymbol{\beta}$  is a vector of fixed effects,  $\mathbf{u}$  is a vector of random effects,  $\mathbf{X}$  and  $\mathbf{Z}$  are known incidence matrices relate  $\mathbf{y}$  to the vectors  $\boldsymbol{\beta}$  and  $\mathbf{u}$ , respectively, and  $\boldsymbol{\varepsilon}$  is a vector of residual terms. Generally, it is assumed that  $\mathbf{u}$  and  $\boldsymbol{\varepsilon}$  are independent from each other and normally distributed with zero-mean vectors and variance–covariance matrices  $\mathbf{G}$  and  $\boldsymbol{\Sigma}$ , respectively.

As mentioned before, in animal breeding a central goal refers to the prediction of random effects (breeding values). In linear (Gaussian) models as in (6) such predictions are given by the conditional expectation of  $\mathbf{u}$  given the data, i.e.,  $E[\mathbf{u}|\mathbf{y}]$ . Given the model specifications above, the joint distribution of  $\mathbf{y}$  and  $\mathbf{u}$  is:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{V} & \mathbf{Z}\mathbf{G} \\ \mathbf{G}\mathbf{Z}' & \mathbf{G} \end{bmatrix} \right),$$

where  $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \boldsymbol{\Sigma}$ .

From the properties of multivariate normal distributions,  $E[\mathbf{u}|\mathbf{y}]$  is given by:

$$E[\mathbf{u}|\mathbf{y}] = E[\mathbf{u}] + \text{Cov}[\mathbf{u}, \mathbf{y}'] \text{Var}^{-1}[\mathbf{y}](\mathbf{y} - E[\mathbf{y}]),$$

such that in this case:

$$\begin{aligned} E[\mathbf{u}|\mathbf{y}] &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{G}\mathbf{Z}'(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \boldsymbol{\Sigma})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

This expression, however, depends on the fixed effects values  $\boldsymbol{\beta}$ , which also need to be inferred from the data. The fixed effects are then typically replaced by their estimates, such that predictions are made based on the following expression:

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

To estimate the fixed effects  $\boldsymbol{\beta}$ , all random effects in Model (6) can be combined into a single vector,

$\xi = \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ , such that the following fixed effects model is obtained:  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \xi$ . It is shown that the expectation of the  $\xi$  term is  $E[\xi] = E[\mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}] = \mathbf{Z}E[\mathbf{u}] + E[\boldsymbol{\varepsilon}] = \mathbf{0}$ , and that its variance is  $\text{Var}[\xi] = \text{Var}[\mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}] = \mathbf{Z}\text{Var}[\mathbf{u}]\mathbf{Z}' + \text{Var}[\boldsymbol{\varepsilon}] = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \boldsymbol{\Sigma} = \mathbf{V}$ . Under these settings, the distribution of  $\mathbf{y}$  is multivariate normal with mean vector  $\mathbf{X}\boldsymbol{\beta}$  and covariance matrix  $\mathbf{V}$ , i.e.,  $\mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ , and the maximum likelihood estimator of  $\boldsymbol{\beta}$  can be shown to be:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

which is distributed as  $\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1})$ . If the design matrix  $\mathbf{X}$  is not full column rank, a generalized inverse of  $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$  must be used to obtain a solution  $\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$  of the system, from which estimable functions  $\boldsymbol{\theta} = \mathbf{L}\boldsymbol{\beta}$  are estimated as  $\hat{\boldsymbol{\theta}} = \mathbf{L}\boldsymbol{\beta}^0$ .

The solutions  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{u}}$  discussed before require  $\mathbf{V}^{-1}$ . As  $\mathbf{V}$  can be of huge dimensions, especially in animal breeding applications, its inverse is generally computationally demanding if not unfeasible. However, Henderson [18] presented the mixed model equations (MME) to estimate  $\boldsymbol{\beta}$  and  $\mathbf{u}$  simultaneously, without the need for computing  $\mathbf{V}^{-1}$ . The MME were derived by maximizing (for  $\boldsymbol{\beta}$  and  $\mathbf{u}$ ) the joint density of  $\mathbf{y}$  and  $\mathbf{u}$ , expressed as:

$$p(\mathbf{y}, \mathbf{u}) \propto |\boldsymbol{\Sigma}|^{-1/2} |\mathbf{G}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) - \frac{1}{2} \mathbf{u}' \mathbf{G}^{-1} \mathbf{u} \right\}.$$

The logarithm of this function is:

$$\begin{aligned} \ell = \log[p(\mathbf{y}, \mathbf{u})] &\propto |\boldsymbol{\Sigma}| + |\mathbf{G}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}' \mathbf{G}^{-1} \mathbf{u} \\ &= |\boldsymbol{\Sigma}| + |\mathbf{G}| + \mathbf{y}' \boldsymbol{\Sigma}^{-1} \mathbf{y} - 2\mathbf{y}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\mathbf{u} \\ &\quad + \boldsymbol{\beta}' \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}' \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\mathbf{u} \\ &\quad + \mathbf{u}' \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\mathbf{u} + \mathbf{u}' \mathbf{G}^{-1} \mathbf{u}. \end{aligned}$$

The derivatives regarding  $\boldsymbol{\beta}$  and  $\mathbf{u}$  are:

$$\begin{bmatrix} \frac{\partial \ell}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell}{\partial \mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y} - \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\hat{\mathbf{u}} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{y} - \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\hat{\mathbf{u}} - \mathbf{G}^{-1} \hat{\mathbf{u}} \end{bmatrix}.$$

Equating them to zero gives the following system:

$$\begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\hat{\mathbf{u}} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\hat{\mathbf{u}} + \mathbf{G}^{-1} \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{y} \end{bmatrix},$$

which can be expressed as:

$$\begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{y} \end{bmatrix},$$

known as the mixed model equations (MME).

Using the second part of the MME,

$$\mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1}) \hat{\mathbf{u}} = \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{y},$$

such that

$$\hat{\mathbf{u}} = (\mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

It can be shown that this expression is equivalent to  $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \boldsymbol{\Sigma})^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$  and, more importantly, that  $\hat{\mathbf{u}}$  is the best linear unbiased predictor (BLUP) of  $\mathbf{u}$ . Using this result into the first part of the MME,

$$\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}\hat{\mathbf{u}} = \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y}$$

$$\begin{aligned} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}(\mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y} \end{aligned}$$

$$\hat{\boldsymbol{\beta}} = \{ \mathbf{X}' [\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{Z}(\mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}' \boldsymbol{\Sigma}^{-1}] \mathbf{X} \}^{-1} \mathbf{X}' [\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{Z}(\mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}' \boldsymbol{\Sigma}^{-1}] \mathbf{y}.$$

Similarly, it is shown that this expression is equivalent to  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ , which is the best linear unbiased estimator (BLUE) of  $\boldsymbol{\beta}$ .

It is important to note that  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{u}}$  require knowledge of  $\mathbf{G}$  and  $\boldsymbol{\Sigma}$ , or at least some function of them. As these matrices are rarely known, the practical approach is to replace  $\mathbf{G}$  and  $\boldsymbol{\Sigma}$  by some sort of point estimates  $\hat{\mathbf{G}}$  and  $\hat{\boldsymbol{\Sigma}}$  into the MME.

Many methods have been proposed to estimate variance components in mixed-effects models. The simplest is the analysis of variance (ANOVA) method, which works well for simple models (such as a one-way structure) or balanced data (such as data from designed experiments with no missing data), but they are not indicated for more complex models and data structures such as those generally found in the animal breeding context.

Alternative methods proposed for estimating variance components in more complex scenarios include

the expected mean squares approach of Henderson [19] and the minimum norm quadratic unbiased estimation [20]. However, maximum likelihood-based methods are currently the most popular (see, for example, [21]), especially the restricted (or residual) maximum likelihood (REML) approach [22], which attempts to correct for the well-known bias in the classical maximum likelihood (ML) estimation of variance components. Additional literature on variance component estimation and mixed model methodology can be found, for example, in [23–27].

### The Animal Model

The advent of mixed-effects models has undoubtedly revolutionized the animal breeding field, and today they are widely used in the genetic improvement of many livestock and companion animal species. In this subsection some of the applications of mixed models for the genetic evaluation of populations using phenotypic and pedigree information will be presented. In the following section applications incorporating molecular marker information will be discussed as well.

As a first application of mixed models in animal breeding, the so-called “animal model” is considered here for the specific situation of a single trait and a single phenotypic observation (including missing values) per animal. The animal model can be described as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon},$$

where  $\mathbf{y}$  is an  $(n \times 1)$  vector of observations (phenotypic scores),  $\boldsymbol{\beta}$  is a  $(p \times 1)$  vector of fixed effects (e.g., herd-year-season effects in cattle evaluations), and  $\boldsymbol{\varepsilon}$  represents residual effects, assumed  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$  as before. In most applications of animal models, however, residuals are assumed independent across animals, such that the residual covariance structure can be expressed as  $\mathbf{R} = \mathbf{I}\sigma_{\varepsilon}^2$ , where  $\mathbf{I}$  is an identity matrix of appropriate order, and  $\sigma_{\varepsilon}^2$  is the residual variance. In the case of animal models, the random effects  $\mathbf{u}$  represent the breeding values, i.e.,  $\mathbf{u} = \mathbf{a}$ , assumed to be  $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G})$ . The vector  $\mathbf{a}$ , of dimension  $(q \times 1)$ , may include breeding values of all animals with record or in the pedigree file, such that  $q$  is generally bigger than  $n$ .

The matrix  $\mathbf{G}$ , which in this case describes the covariances among the breeding values, follows from standard results for the covariances between relatives.

It is seen that the additive genetic covariance between two relatives  $i$  and  $i'$  is given by  $2\theta_{ii'}\sigma_a^2$ , where  $\theta_{ii'}$  is the coefficient of coancestry between individuals  $i$  and  $i'$ , and  $\sigma_a^2$  is the additive genetic variance in the base population [28]. Hence, under the animal model,  $\mathbf{G} = \mathbf{A}\sigma_a^2$ , where  $\mathbf{A}$  is the “additive genetic (or numerator) relationship matrix,” having elements given by  $a_{ii'} = 2\theta_{ii'}$ .

As mentioned earlier, in animal breeding the usual main interest is prediction of breeding values – for selection of superior individuals – and on estimation of variance components. The fixed effects are, in some sense, nuisance factors with no central interest in terms of inferences, but which need to be taken into account (i.e., they need to be corrected for when inferring breeding values).

Because under the animal model  $\mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_a^{-2}$  and  $\mathbf{R}^{-1} = \mathbf{I}\sigma_{\varepsilon}^{-2}$ , the mixed model equations reduce to:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

where  $\lambda = \frac{\sigma_{\varepsilon}^2}{\sigma_a^2} = \frac{1 - h^2}{h^2}$ , such that:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}.$$

It is worth mentioning that  $\mathbf{A}^{-1}$  can be obtained directly from the pedigree, without setting up  $\mathbf{A}$  [29, 30], which is computationally very convenient.

Conditional on the variance components ratio  $\lambda$ , the BLUP of the breeding values are given then by  $\hat{\mathbf{a}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ , which are the estimated breeding values (EBV). Alternatively, some breeders' associations express their results as predicted transmitting abilities (PTA) or expected progeny differences (EPD), which are equal to half the EBV, representing the portion of an animal's breeding values that is passed to its offspring.

The amount of information contained in an animal's genetic evaluation depends on the availability of its own record, and of phenotypic information from its relatives (including how many and how closely related to it). As a measure of amount of information in livestock genetic evaluations, EBV are typically reported with their associated accuracies, i.e. the correlation between true and estimated breeding values,  $r_i = r_{\hat{a}_i, a_i}$ . Instead of accuracy, some livestock species

genetic evaluations use reliability, which is the accuracy squared ( $r_i^2$ ).

The calculation of  $r_i$  requires the diagonal elements of the inverse of the MME coefficient matrix, represented as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta a} \\ \mathbf{C}^{a\beta} & \mathbf{C}^{aa} \end{bmatrix}.$$

It is shown that the prediction error variance (PEV) of EBV  $\hat{a}_i$  is given by:

$$\text{PEV} = \text{Var}(\hat{a}_i - a_i) = c_i^{aa} \sigma_a^2,$$

where  $c_i^{aa}$  is the  $i$ th diagonal element of  $\mathbf{C}^{aa}$ , relative to animal  $i$ . The PEV can be interpreted as the fraction of additive genetic variance not accounted for by the prediction. Therefore, PEV can also be expressed as:

$$\text{PEV} = (1 - r_i^2) \sigma_a^2,$$

such that  $c_i^{aa} \sigma_a^2 = (1 - r_i^2) \sigma_a^2$ , from which the reliability is obtained as  $r_i^2 = 1 - c_i^{aa} \sigma_a^2 / \sigma_a^2 = 1 - \lambda c_i^{aa}$ .

### Extensions and Variations of the Animal Model

The animal model discussed above can be extended also to multiple (correlated) traits [31, 32]. For instance, consider as an example the analysis of  $k$  traits, in which the model for each trait is expressed as:

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{a}_j + \boldsymbol{\varepsilon}_j,$$

where  $\mathbf{y}_j$ ,  $\mathbf{X}_j$ ,  $\boldsymbol{\beta}_j$ ,  $\mathbf{Z}_j$ ,  $\mathbf{a}_j$ , and  $\boldsymbol{\varepsilon}_j$  are defined as before, but here have an additional index to indicate the trait ( $j = 1, 2, \dots, k$ ).

For a joint analysis of the  $k$  traits, the single trait models can be combined as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon},$$

where  $\mathbf{y} = [\mathbf{y}_1' \ \mathbf{y}_2' \ \dots \ \mathbf{y}_k']'$ ,  $\boldsymbol{\beta} = [\boldsymbol{\beta}_1' \ \boldsymbol{\beta}_2' \ \dots \ \boldsymbol{\beta}_k']'$ ,  $\mathbf{a} = [\mathbf{a}_1' \ \mathbf{a}_2' \ \dots \ \mathbf{a}_k']'$ , and  $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1' \ \boldsymbol{\varepsilon}_2' \ \dots \ \boldsymbol{\varepsilon}_k']'$ , and the design matrices in this case are:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_k \end{bmatrix} \text{ and } \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Z}_k \end{bmatrix}.$$

$$\text{It is assumed that } \text{Var} \begin{bmatrix} \mathbf{a} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \otimes \mathbf{I} \end{bmatrix},$$

$$\text{where } \mathbf{G} = \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2} & \dots & \sigma_{a_1 a_k} \\ \sigma_{a_1 a_2} & \sigma_{a_2}^2 & \dots & \sigma_{a_2 a_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_1 a_k} & \sigma_{a_2 a_k} & \dots & \sigma_{a_k}^2 \end{bmatrix} \text{ and}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_1 \varepsilon_2} & \dots & \sigma_{\varepsilon_1 \varepsilon_k} \\ \sigma_{\varepsilon_1 \varepsilon_2} & \sigma_{\varepsilon_2}^2 & \dots & \sigma_{\varepsilon_2 \varepsilon_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon_1 \varepsilon_k} & \sigma_{\varepsilon_2 \varepsilon_k} & \dots & \sigma_{\varepsilon_k}^2 \end{bmatrix} \text{ are the genetic and}$$

residual variance–covariance matrices, respectively,  $\mathbf{A}$  and  $\mathbf{I}$  are the numerator relationship matrix and an identity matrix, and  $\otimes$  represents the direct (Kronecker) product.

The MME for multi-trait analyses are of the same form as before, i.e.,

$$\begin{bmatrix} \mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{X} & \mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z} \\ \mathbf{Z}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{X} & \mathbf{Z}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{Z} + \mathbf{G}^{-1} \otimes \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y} \\ \mathbf{Z}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y} \end{bmatrix},$$

from which the BLUEs and BLUPs of  $\boldsymbol{\beta}$  and  $\mathbf{a}$  can be obtained, respectively.

The dimensionality of such multi-trait MME, however, can become a hurdle for solving it when more than two or three traits are considered. An alternative for the analysis of multiple traits is to use a canonical transformation of the traits [33–35], which consists of transforming the vectors of correlated traits into a new vector of uncorrelated variables. In such case, each transformed variable can be analyzed independently using standard single trait models, and subsequently the estimated breeding values are transformed back to the original scale of measurement.

Some other interesting applications of mixed models in animal breeding involve multiple random effects, as in the cases of repeated measurements of the same trait or traits with maternal effects. For the analysis of repeated measurements, as discussed in section “Selection Index” (Model 4), environmental effects can be partitioned into permanent and temporary effects. In this case, the mixed model, usually called “repeatability model,” can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + \boldsymbol{\varepsilon},$$

where all terms are as previously defined for a single trait animal model, and  $\mathbf{p}$  is the vector of permanent

environmental effects, with each level pertaining to a common effect to all observations of each animal, and  $\mathbf{W}$  is a known incidence matrix relating  $\mathbf{y}$  to the vector  $\mathbf{p}$ .

It is often assumed that  $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$ ,  $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I}\sigma_p^2)$ , and  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$  are independent from each other. Under these assumptions, the MME becomes:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda_a\mathbf{A}^{-1} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W} + \lambda_p\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix},$$

where  $\lambda_a = \sigma_e^2/\sigma_a^2$  and  $\lambda_p = \sigma_e^2/\sigma_p^2$ .

There are some traits of interest in livestock, such as weaning weight in beef cattle, in which progeny performance is affected by the dam's ability to affect the calf's environment, such as in the form of nourishment through her milk production, the quantity and quality of which is in part genetically determined. In some cases, there can also be a paternally provided environmental component. In such cases, parents contribute to the performance of their progeny not only through the genes passed to the progeny (the "direct genetic effects") but also through their ability to provide a suitable environment (the "indirect genetic effects").

Here maternally influenced traits are considered, for which the mixed model can be written as [36]:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{K}\mathbf{m} + \mathbf{W}\mathbf{p} + \boldsymbol{\varepsilon},$$

where all terms are as before, except that the model now includes a vector  $\mathbf{m}$  of random maternal genetic effects, and a vector  $\mathbf{p}$  of random permanent environmental effects, with  $\mathbf{K}$  and  $\mathbf{W}$  as their respective incidence matrices. It is assumed that  $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$ ,  $\mathbf{m} \sim N(\mathbf{0}, \mathbf{A}\sigma_m^2)$ ,  $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I}\sigma_p^2)$ , and  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ , and quite often a covariance structure between direct and maternal additive genetic effects is considered, assumed equal to  $\mathbf{A}\sigma_{a,m}$ .

Some other variations of the animal model, which are computationally convenient, include the "sire model" and the "reduced animal model" [37]. In the sire models, only sires are evaluated, using progeny records under the assumption of randomly selected mates. In the reduced animal model, instead of having equations set up for every animal (i.e., parents and progeny), it allows equations to be set up only for parents in the MME, making the dimensions of

the system greatly reduced. The breeding values of the parents are estimated directly from the MME, and the progeny breeding values are then inferred by back solving from the predicted parental breeding values.

As a final note regarding the use of mixed models in animal breeding, it is important to mention that solving the MME does not necessarily require the inversion of the coefficient matrix  $\mathbf{C}$ . More computationally convenient alternatives for solving high dimensional systems of linear equations include methods based on iteration on the MME, such as the Jacobi or Gauss-Seidel iteration [38], and the "iteration on the data" strategy [39], which is a commonly used methodology in national genetic evaluations involving millions of records.

## Marker-Assisted Selection

### Introduction

The advent of molecular markers has created opportunities for a better understanding of genetic inheritance and for developing novel strategies for genetic improvement in agriculture. Molecular markers are used, for example, to study quantitative trait loci (QTL), which are defined as chromosomal regions contributing to variation in phenotypic traits. The location and effects of QTL can be inferred by combining information from marker genotypes and phenotypic scores of individuals and by exploring genetic linkage [40–43] and linkage disequilibrium [44, 45] information between marker loci and QTL, such as in experimental or mapping populations (e.g., backcross or  $F_2$ , or granddaughter designs) or in complex pedigrees in outbred populations. Information on markers associated with QTL can be used to enhance prediction of genetic merit of animals [46]. This is especially useful for low heritability traits, traits that are expensive or difficult to measure, or traits expressed in only one sex [47].

### Classical Approaches with Few Markers

The application of molecular information for genetic improvement of animals and plants, or marker-assisted selection (MAS), requires that candidate-for-selection individuals are genotyped for specific markers. For MAS purposes, there are three types of genetic markers, and for each type there are specific statistical



approaches for incorporating their information into selection programs [47]. A first type of marker refers to situations in which the functional polymorphism itself can be genotyped. These markers are called “direct markers,” as they indicate exactly the genotype an animal has at specific causative loci.

A second type of marker refers to those that are in population-wide linkage disequilibrium (LD) with the causative or functional mutations. In such cases, although the marker genotype of an animal does not unambiguously indicate the genotype at a specific functional locus, it still provide information regarding how likely an animal carries a specific allele or genotype at such a locus. Finally, a third kind of molecular marker refers to those loci that are in population linkage equilibrium with the functional mutations, which are often called “indirect markers.” In such cases, although the marker information on a single animal in a population does not provide any information regarding the genetic merit of that animal, it still can be useful in exploring family (pedigree) structure when genotyped animals are related to each other.

While direct markers are the simplest and most efficient in MAS programs, their identification is much more difficult and generally involves a prescreening step using QTL mapping methods to identify promising chromosomal regions, followed by fine mapping (often using functional and positional candidate gene strategies), followed by validation (using some strategy such as a knock-out approach). On the other extreme, indirect markers are extensively available for most livestock species, but their use in MAS is more complex and the results are generally modest.

Statistical models to incorporate direct and/or LD markers in the genetic evaluations of animals are relatively straightforward. For example, a marker can be included into an animal model context with the following specification:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}^* + \mathbf{M}\mathbf{g} + \boldsymbol{\varepsilon},$$

where all terms are as defined before, except that  $\mathbf{a}^* \sim N(\mathbf{0}, \mathbf{A}\sigma_{a^*}^2)$  represents now the random additive (nonmarked) polygenic effects, and  $\mathbf{g}$  and  $\mathbf{M}$  are the (fixed) QTL effects and an incidence matrix, respectively. In the case of direct markers, the matrix  $\mathbf{M}$  represents the marker genotypes and is obtained directly from the genotyping of animals. In the case of

LD markers, the incidence matrix  $\mathbf{M}$  will represent genotype probabilities at each QTL locus, which can be derived using segregation analysis. The overall genetic merits of the animals are then given by the sum of their  $\mathbf{a}^*$  and  $\mathbf{g}$  components. Other strategies for combining the infinitesimal and the QTL components to increase long-term genetic gain have also been proposed (e.g., [48–50]); a review of MAS strategies can be found, for example, in [47].

In the case of indirect markers, however, the within-family LD between QTL and linked markers must be explored. One approach is to determine the marker effects or the marker-QTL linkage phases separately for each family. Alternatively, more general MAS models have been proposed to incorporate marker data in genetic evaluations for complex pedigrees [13, 51], which can be represented as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}^* + \mathbf{M}\mathbf{q} + \boldsymbol{\varepsilon},$$

where the terms are as before, but here the QTL effects  $\mathbf{q}$  are assumed random and normally distributed, such that:

$$\begin{bmatrix} \mathbf{a}^* \\ \mathbf{q} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{A}\sigma_{a^*}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_\lambda\sigma_q^2 \end{bmatrix}\right),$$

where  $\mathbf{G}_\lambda$  is the gametic relationship matrix for the QTL, and  $\sigma_q^2$  is the additive variance of the QTL allelic effects. The gametic relationship matrix gives the probabilities of identity between each of the two alleles in each individual, and it can be derived based on the QTL position  $\lambda$  and the marker information.

## Genomic Selection

As most quantitative traits are influenced by many genes, tracking a small number of them using molecular markers (as in the MAS approaches discussed above) will explain only a small fraction of the total genetic variance. Moreover, individual genes are likely to have small effects and so a large amount of data is needed to accurately estimate their effects [52]. Genome-wide Marker-Assisted Selection (GWMAS), or simply Genomic Selection (GS), on the other hand, makes use of a very dense set of markers covering the entire genome, which potentially explain all genetic variance. In addition, given the LD between the dense markers and the QTL, estimated marker effects pertain across the population [53].

Meuwissen et al. [54] were the first to propose GS and suggested a model that can be described as:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^p \mathbf{M}_j \mathbf{q}_j + \boldsymbol{\varepsilon},$$

where  $\mu$  is the overall mean,  $\mathbf{q}_j$  represents the genetic effects of marked genes ( $j=1, 2, \dots, p$ ), and  $\mathbf{M}_j$  represents the design matrices (genotypes) relative to a large number ( $p$ ) of biallelic markers (e.g., SNP loci), which present different levels of LD with QTL affecting the phenotypic trait of interest ( $\mathbf{y}$ ). Here it is assumed that the QTL affecting the trait act additively, and that  $\mathbf{q}_j$  refers to per-allele effects; nonadditive effects as well as effects relative to nonmarked QTL are lumped together into the residual term of the model.

Fitting such GS model using standard regression approaches is not trivial, as the number  $p$  of markers (and so the number of genetic effects to be estimated) may easily exceed the number  $n$  of individuals available. The “large  $p$  small  $n$  paradigm” is central in many applications of genomic technologies, including expression profiling and association analysis, and various statistical strategies have been proposed in the literature to overcome this problem, such as dimension-reduction techniques, stepwise fitting procedures, ridge regression [55], and least absolute selection and shrinkage operator – LASSO [56].

Specifically in GS, hierarchical modeling has become the methodology of choice, due to its flexibility and good statistical properties. Within this approach, the genetic effects  $\mathbf{q}_j$  are assumed random effects and distributed according to some prespecified distribution [54]. For example,  $\mathbf{q}_j$  may be assumed normally distributed with mean 0 and variance  $\sigma_j^2$ , and the hierarchy can be extended by assuming a prior distribution for the variances  $\sigma_j^2$  [54, 57–59]. Alternative distributions can be adopted for the  $\mathbf{q}_j$ , such as double exponential or mixture distributions including a mass point at zero. It is interesting to note the connection between the ridge regression approach and a Bayesian model with normal priors with common variances  $\sigma_j^2 = \sigma_0^2$ , as well as the LASSO methodology and a Bayesian model with double exponential priors for the genetic effects [60].

The potential of GS to accelerate genetic progress has been demonstrated through many simulation studies (e.g., [54, 61, 62]), and confirmed by some real data

applications. The first use of GS using thousands of markers in livestock has been in dairy cattle [63, 64], followed by some breeds of beef cattle and more recently in poultry. Table 1 shows some encouraging results on dairy cattle obtained by the USDA.

Future Directions

As shown here, the mixed model methodology is extremely flexible and can be used in a wide variety of applications. Other extensions of the methods discussed here include models with nonadditive genetic effects (e.g., [65, 66]), mixed models for the analysis of non-Gaussian traits such as binary and categorical (e.g., [67, 68]) or counting data (e.g., [69]), robust models [70, 71], survival traits [72], nonlinear models to study, for example, growth curves (e.g., [73, 74]), among others. However, such models can get extremely complex and asymptotic statistical methods are generally required. Alternatively, Bayesian analysis employing Markov Chain Monte Carlo (MCMC) methods can be used, given their exceptional flexibility and the possibility of incorporating prior information regarding the model parameters [75].

Animal Breeding, Foundations of. Table 1 Comparison of April 2010 genomic and traditional evaluations for bulls with an AI status of active or foreign

| Trait                   | Average reliability (%) |             |            |
|-------------------------|-------------------------|-------------|------------|
|                         | Genomic                 | Traditional | Difference |
| Net merit               | 87                      | 81          | +6         |
| Milk yield              | 93                      | 91          | +2         |
| Fat yield               | 93                      | 91          | +2         |
| Protein yield           | 93                      | 91          | +2         |
| Productive life         | 81                      | 71          | +9         |
| Somatic cell score      | 88                      | 83          | +5         |
| Daughter pregnancy rate | 79                      | 69          | +10        |
| Final score             | 89                      | 85          | +4         |
| Sire calving ease       | 90                      | 84          | +6         |
| Daughter calving ease   | 80                      | 67          | +13        |

Source: AIPL – USDA; <http://www.aipl.arsusda.gov/>

Bayesian analysis has been increasingly used in many applications of genetics and animal breeding, and for a review the reader can refer, for example, to [76–78]. A comprehensive treatment of Bayesian MCMC approaches in animal breeding is presented in [23].

Bayesian hierarchical modeling has also been extensively used in genomic selection [54, 79–81]. In addition, nonparametric and semiparametric methods, and machine learning techniques based on artificial intelligence have been proposed and used for the analysis of high density marker panels in the context of animal breeding, such as in [82–86]. Moreover, some other recent methods aim to combine all available phenotypic, pedigree, and genomic information for prediction of genetic merit of animals [87].

As indicated in the beginning of this chapter, the genetic improvement observed in many livestock and companion animal species is truly remarkable. Most of this genetic progress has been accomplished through selection, using the methods discussed here. Two technological and methodological developments however must be mentioned as turning points in the genetic trends observed in some species; these are the advent of artificial insemination and the mixed models. Seemingly, the development of high density SNP panels and, more recently, next generation sequencing technologies and their application in genomic selection strategies promise to be the next turning point. This new era for animal breeding and genetics will require a different profile of animal breeders, requiring not only knowledge of population and quantitative genetics, classical statistical and computational methods, but also some more modern statistical and computational methods based on hierarchical modeling [23, 88], non- and semiparametric methods, and machine learning techniques [60, 89]. It is indeed a very exciting time to work in animal breeding!

## Bibliography

### Primary Literature

- Lush JL (1994) The genetics of populations. (Prepared for publication by Chapman AB, Shrode RR, with an addendum by Crow JF) Special report 94, College of Agriculture, Iowa State University, Ames
- Bulmer MG (1985) The mathematical theory of quantitative genetics. Clarendon, Oxford
- Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longmans Green, Harlow
- Lynch M, Walsh B (1998) Genetic analysis of quantitative traits. Sinauer Associates, Sunderland
- Hill WG (1969) On the theory of artificial selection in finite populations. *Genet Res* 13:143–163
- Havenstein B, Ferket PR, Qureshi MA (2003) Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. *Poult Sci* 82:1509–1518
- Bourdon RM (2000) Understanding animal breeding, 2nd edn. Prentice Hall, Upper Saddle River
- Crow J, Kimura M (1970) An introduction to populations genetics theory. Harper and Row, New York
- Shook GE (2006) Major advances in determining appropriate selection goals. *J Dairy Sci* 89(4):1349–1361
- Henderson CR (1949) Estimation of changes in herd environment. *J Dairy Sci* 32:709
- Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31:423–447
- Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph
- Fernando RL, Grossman M (1989) Marker-assisted selection using best linear unbiased prediction. *Genet Sel Evol* 21:467–477
- Yu J et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* 8:625–637
- Rosa GJM, Steibel JP, Tempelman RJ (2005) Reassessing design and analysis of two-color microarray experiments using mixed effects models. *Comp Funct Genomics* 6:123–131
- Steibel JP, Poletto R, Coussens PM, Rosa GJM (2009) A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data. *Genomics* 94:146–152
- Henderson CR (1950) Estimation of genetic parameters. *Ann Math Stat* 21:309
- Henderson CR (1953) Estimation of variance and covariance components. *Biometrics* 9:226
- Rao CR (1971) Estimation of variance and covariance components MINQUE theory. *J Multivar Anal* 1:257–275
- Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc* 72(358):320–338
- Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3):545–554
- Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York
- Littell RC, Milliken GA, Stroup WW, Wolfinger RD (2006) SAS system for mixed models, 2nd edn. SAS Institute, Cary

25. Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-Plus. Springer, New York
26. Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York
27. Verbeke G, Molenberghs G (1997) Linear mixed models in practice: A SAS-oriented approach, Lecture Notes in Statistics 126. Springer, New York
28. Wright S (1921) Systems of mating I. The biometric relations between parents and offspring. *Genetics* 6:111–123
29. Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32:69–83
30. Quaas RL (1976) Computing the diagonal elements of a large numerator relationship matrix. *Biometrics* 32:949–953
31. Henderson CR, Quaas RL (1976) Multiple trait evaluation using relatives' records. *J Anim Sci* 43:1188–1197
32. Schaeffer LR (1984) Sire and cow evaluation under multiple trait models. *J Dairy Sci* 67:1567–1580
33. Thompson R (1977) Estimation of quantitative genetic parameters. In: Pollak E, Kempthorne O, Bailey TB (eds) Proceedings of the international conference on quantitative genetics, Iowa State University Press, Ames, pp. 639–657
34. Meyer K (1985) Maximum-likelihood estimation of variance-components for a multivariate mixed model with equal design matrices. *Biometrics* 41:153
35. Ducrocq V, Besbes B (1993) Solution of multiple trait animal models with missing data on some traits. *J Anim Breed Genet* 110:81–92
36. Quaas RL, Pollak EJ (1981) Modified equations for sire models with groups. *J Dairy Sci* 64:1868–1872
37. Quaas RL, Pollak EJ (1980) Mixed model methodology for farm and ranch beef cattle testing programs. *J Anim Sci* 51:1277–1287
38. Misztal I, Gianola D (1988) Indirect solution of mixed model equations. *J Dairy Sci* 77(Suppl 2):99–106
39. Schaeffer LR, Kennedy BW (1986) Computing solutions to mixed model equations. In: 3rd World congress on genetic applied livestock production, Lincoln, Nebraska, 16–22 July 1986, vol XII. pp 382–393
40. Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185–199
41. Haley CS, Knott SA (1992) A simple regression method to for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69:315–324
42. Haley CS, Knott SA, Elsen J-M (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* 136:1195–1207
43. Pérez-Enciso M, Misztal I (2004) Qxpack: a versatile mixed model application for genetical genomics and QTL analyses. *Bioinformatics* 20(16):2792–2798
44. Meuwissen THE, Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155:421–430
45. Pérez-Enciso M (2003) Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics* 163:1497–1510
46. Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756
47. Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. *Nat Rev Genet* 3(1):22–32
48. Dekkers JCM, van Arendonk JAM (1998) Optimizing selection for quantitative traits with information on an identified locus in outbred populations. *Genet Res* 71(3):257–275
49. Manfredi E, Barbieri M, Fournet F, Elsen JM (1998) A dynamic deterministic model to evaluate breeding strategies under mixed inheritance. *Genet Sel Evol* 30:127–148
50. Chakraborty R, Moreau L, Dekkers JCM (2002) A method to optimize selection on multiple identified quantitative trait loci. *Genet Sel Evol* 34(2):145–170
51. Goddard ME (1992) A mixed model for analyses of data on multiple genetic-markers. *Theor Appl Genet* 83:878–886
52. Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* 124(6):323–330
53. Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet* 123:218–223
54. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
55. Whittaker JC, Thompson R, Visscher PM (2000) Marker-assisted selection using ridge regression. *Genet Res* 75:249–252
56. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J R Stat Soc B* 58:267–288
57. Gianola D, Perez-Enciso M, Toro MA (2003) On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* 163:347–365
58. Xu S (2003) Estimating polygenic effects using markers of the entire genome. *Genetics* 163(2):789–801
59. ter Braak CJF, Boer MP, Bink MCAM (2005) Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* 170(3):1435–1438
60. Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: data mining, inference, and predictions. Springer, New York
61. Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J Anim Breed Genet* 124:362–368
62. Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J Anim Breed Genet* 124:342–355
63. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor J, Schenkel FS (2009) Reliability of genomic predictions for North American dairy bulls. *J Dairy Sci* 92:16–24

64. Weigel KA, de Los Campos G, González-Recio O, Naya H, Wu XL, Long N, Long N, Rosa GJM, Gianola D (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci* 92:5248–5257
65. Henderson CR (1985) Best linear unbiased prediction of non-additive genetic merits in non-inbred populations. *J Anim Sci* 60:111–117
66. Hoeschele I, VanRaden PM (1991) Rapid inverse of dominance relationship matrices for noninbred populations by including sire and dam subclass effects. *J Dairy Sci* 74:557–569
67. Gianola D (1982) Theory and analysis of threshold characters. *J Anim Sci* 54:1079–1096
68. Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical-data with a threshold-model. *Genet Sel Evol* 15(2): 201–223
69. Tempelman RJ, Gianola D (1996) A mixed effects model for overdispersed count data in animal breeding. *Biometrics* 52:265–279
70. Strandén I, Gianola D (1998) Attenuating effects of preferential treatment with student-t mixed linear models: a simulation study. *Genet Sel Evol* 31:25–42
71. Rosa GJM, Padovani CR, Gianola D (2003) Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biom J* 45(5):573–590
72. Ducrocq V, Casella G (1996) A Bayesian analysis of mixed survival models. *Genet Sel Evol* 28(6):505–529
73. Varona L (1997) Multiple trait genetic analysis of underlying biological variables of production functions. *Livest Prod Sci* 47:201–209
74. Forni S, Piles M, Blasco A et al (2009) Comparison of different nonlinear functions to describe Nelore cattle growth. *J Anim Sci* 87(2):496–506
75. Gianola D, Fernando RL (1986) Bayesian methods in animal breeding theory. *J Anim Sci* 63:217–244
76. Shoemaker JS, Painter IS, Weir BS (1999) Bayesian statistics in genetics - a guide for the uninitiated. *Trends Genet* 15:354–358
77. Blasco A (2001) The Bayesian controversy in animal breeding. *J Anim Sci* 79(8):2023–2046
78. Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nat Rev Genet* 5:251–261
79. Yi N, Xu S (2008) Bayesian Lasso for quantitative trait loci mapping. *Genetics* 179:1045–1055
80. Gianola D, de Los Campos G, Hill WG et al (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183(1):347–363
81. de Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes J (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics* 182:375–385
82. Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173:1761–1776
83. Gianola D, van Kaam JBCHM (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178:2289–2303
84. Long N, Gianola D, Rosa GJM, Weigel KA, Avendaño S (2007) Machine learning procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J Anim Breed Genet* 124(6):377–389
85. González-Recio O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendano S (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* 178(4):2305–2313
86. Campos G, Gianola D, Rosa GJM (2009) The linear model of quantitative genetics is a reproducing kernel Hilbert spaces regression. *J Anim Sci* 87:1883–1887
87. Misztal I, Legarra A, Aguilar I (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J Dairy Sci* 92:4648–4655
88. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall, London
89. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

## Books and Reviews

- Chapman AB (1980) General and quantitative genetics. World animal science series. Elsevier, Amsterdam
- Lange K (1997) Mathematical and statistical methods for genetic analysis. Springer, New York
- Liu BH (1998) Statistical genomics. CRC Press, Boca Raton
- Mrode R (2005) Linear models for the prediction of animal breeding values, 2nd edn. CAB International, New York
- Ott J (1991) Analysis of Human Genetic Linkage. Johns Hopkins, Baltimore
- Sham PC (1998) Statistics in human genetics. Arnold, London
- Van Vleck LD (1993) Selection index and introduction to mixed model methods for genetic improvement of animals. CRC Press, Boca Raton

## Animal Breeding, Modeling in

LAWRENCE R. SCHAEFFER

Department of Animal and Poultry Science, Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada

## Article Outline

Glossary

Definition of the Subject

Introduction



Animal Model  
 Genetic Relationships  
 Genetic Evaluation  
 Single Record per Animal  
 Repeated Records Animal Model  
 Maternal Effects Animal Model  
 Random Regression Animal Model  
 Multiple Trait Models  
 Models for Categorical Data  
 Models for Survival Data  
 Added Complexities  
 Future Directions  
 Bibliography

## Glossary

**Contemporary groups** A group of animals of approximately the same age living in the same environment and being treated by the same management practices during the same interval of time.

**Estimated breeding value** An estimate of the total additive genetic merit of an individual, the effects that are directly passed to offspring.

**Genotypes** The particular set of alleles at all gene loci that influence the phenotypes.

**Inbreeding coefficients** The proportion of alleles at gene loci that are identical due to being inherited from a common ancestor.

**Infinitesimal genetic model** A genetic model that assumes there are an infinite number of gene loci affecting a trait each with a small and equal effect.

**Mixed model equations** Proposed by Henderson in 1949 for the estimation of breeding values and other nongenetic effects from phenotypes.

**Phenotypes** The observable characteristics of an animal that can be measured, scored, or recorded.

## Definition of the Subject

Modeling in animal breeding involves describing the major factors that influence the performance ability or production level of animals in order to predict the genetic merit of future progeny for that ability. Successful modeling depends on good record collection systems, accurate pedigree records, and sophisticated statistical models. Models have evolved over time as computer technology has advanced. Genetic evaluation of dairy bulls began in the early 1930s using simple

daughter averages for milk production in selection index procedures of Lush and his students [1]. Genetic evaluation systems spread to all livestock and to many countries due to Lush. Henderson [2] introduced best linear unbiased prediction (BLUP) around 1950, and this methodology is still widely used in animal breeding except that the models are more detailed and complex. Gianola and others [3, 4] taught animal breeders how to use Bayesian methods which are especially useful for non-normally distributed data.

All methods consider the total additive genetic merit of animals. Additive genetic effects are directly transmitted to progeny, assuming infinitely many gene loci, each with a small and equal effect on the trait of interest. Examples of traits are milk yields of dairy animals, growth rates of meat-producing animals, occurrence of health problems, ability to reproduce, and behavior.

Models to analyze traits differ depending on the nature of the traits being evaluated. Traits can be broadly grouped into production traits, growth traits, reproductive traits, health traits, and survival, and the models for each group are very specialized. In the 1930s, dairy bulls were selected almost solely for the ability of their daughters to produce milk. Over time, the success of that selection pressure for one trait caused negative correlated responses in reproduction potential and disease susceptibility. Genetic evaluation systems consider many traits simultaneously with a primary focus on efficiency of economic production and cost savings.

## Introduction

Modeling in animal breeding began with the expression

$$\mathbf{p} = \mathbf{g} + \mathbf{e},$$

where  $\mathbf{p}$  is the phenotype of an individual,  $\mathbf{g}$  is the genetic source of variation or genotype in  $\mathbf{p}$ , and  $\mathbf{e}$  is the environmental source of variation, presented by Wilhelm Ludwig Johannsen in 1909 [5]. A phenotype is a measurement of an animal's performance for an economically important trait, such as daily milk yield of a dairy cow, goat, or sheep; a growth record on a beef animal, pig, or rabbit; or wool production of a sheep or llama. Current models are extensions of this basic equation.

## Genetic Component

The  $\mathbf{g}$  term of the model is partitioned as

$$\mathbf{g} = \mathbf{g}_a + \mathbf{g}_d + \mathbf{g}_i,$$

where

$\mathbf{g}_a$  are additive genetic effects, the collective effects generated by individual alleles at every loci in the genome.

$\mathbf{g}_d$  are the dominance genetic effects, the collective effects generated by combinations of alleles at every locus in the genome.

$\mathbf{g}_i$  are the many interactions among loci throughout the genome, in a collective sense.

Animal breeders are primarily concerned with additive genetic effects because those effects are passed directly from parents to progeny. Dominance and interaction genetic effects occur depending on the combinations of alleles from the paternal and maternal sides of the pedigree, and cannot be predicted prior to making a mating of specific individuals.

Further,  $\mathbf{g}_a$  is assumed to be the combined effects of an infinite number of loci, each with a small and equal effect, i.e.,

$$\mathbf{g}_a = \sum_{k=1}^{\infty} \mathbf{g}_k.$$

The population is assumed to be very large and mating randomly. This is known as the *infinitesimal*

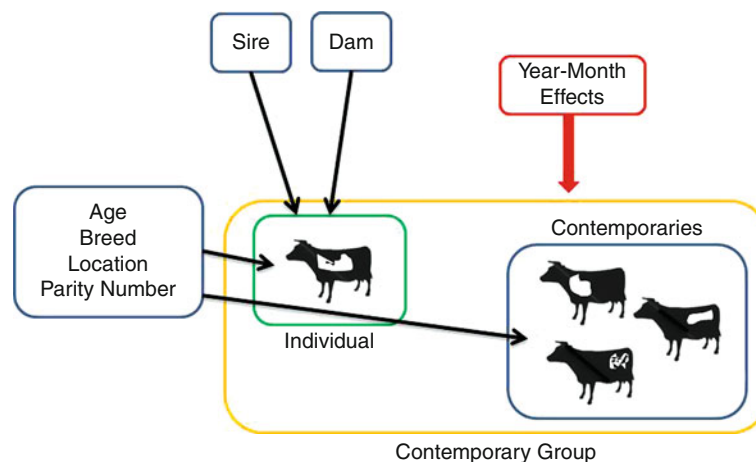
*genetic model* upon which genetic evaluation systems are based. This is a reasonable model if you consider that there are roughly 30,000 loci (this estimate is continually changing), and if one assumes only two alleles at each locus, then there are three possible genotypes per locus, or  $3^{30,000}$  possible genotype outcomes. This number is large enough to be nearly infinite for all practical purposes. The number of possible genotype outcomes is actually much bigger if you allow that there are more than two alleles for many loci.

## Environmental Component

The  $\mathbf{e}$  term can be split into identifiable factors that are known to have effects on the phenotypes. Some effects are shared with other animals in the same location at the same time, and some effects are specific to individuals.

## Animal Model

The common model in animal breeding is called an *animal model*, based on the underlying infinitesimal genetic model, containing factors to account for time trends, contemporary groups, additive genetic effects, residual, and other factors, that depend on the species and trait of interest. Figure 1 depicts a typical animal model in a diagrammatic format. Each individual is affected by the additive genetic merits of its sire and dam, as well as year and month of birth or calving.



Animal Breeding, Modeling in. Figure 1

Diagram of genetic and environmental factors affecting dairy cow milk production

Contemporaries are animals that share the same year and month effects as the individual because they exist in the same space and time. These effects include weather, location, herd owner, and herd management practices of feeding, breeding, and health care. Individual production is affected by breed, age, and parity number, and these effects are common to all animals of the same breed, age, and parity number.

Depending on the species and trait, other factors could be included in the diagram.

Models in animal breeding are conveniently presented in matrix notation. Models consist of three parts which are the equation, the expectations and distributions of random variables, and finally a list of assumptions and limitations. The three parts are briefly described.

### The Equation

A linear statistical model is written generally (in matrix notation) as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where

$\mathbf{y}$  is a vector of observations on the trait of interest from individual animals, either one or more per animal.

$\mathbf{b}$  is a vector of levels of factors that are known to affect the magnitude of the observations. These factors are fixed factors, in that they are constants and not random variables drawn from a particular distribution or population of effects. One factor that should always be included in an animal model is one that accounts for time, either year, or year-season. Seasons may be groups of months, individual months or weeks of the year. This time factor is meant to account for changes in average production over time due to changes in production technology, global warming, nutritional improvements, or financial impacts that affect the entire population. Other factors are effects of gender, breed, age of dam, and diet.

$\mathbf{u}$  is a vector of levels of random factors, which are random samples from a large population of levels of that effect. Animal additive genetic effects are one such random factor contained in  $\mathbf{u}$ . Two random factors that should always be included in an animal model are the animal additive genetic effect,  $\mathbf{a}$ , and

the contemporary group effect,  $\mathbf{c}$ , ( $\mathbf{u}' = (\mathbf{a}' \ \mathbf{c}')$ ). Contemporaries are animals that coexist in the same location and time space, and thus share the same management care and treatment during the time they are observed for a trait. Contemporary groups are commonly called herds, flocks, or tanks, and are nested within year or year-season effects. Contemporary groups are random because there are a large number of them, each contemporary group contains different individuals, and the effects during that one occurrence are random in nature.

$\mathbf{e}$  is a vector of random residual (environmental) effects specific to each observation, that cannot be accounted for by other factors in the model. The residual effects may be samples from different populations having different variances.

### Expectations and Covariances

The model also describes the distributions of the random factors and indicates their expected values or means and covariance structures. That is,

$$\text{Var}(\mathbf{u}) = \mathbf{G}$$

$$\text{Var}(\mathbf{e}) = \mathbf{R}, \text{ and}$$

$$\text{Cov}(\mathbf{u}, \mathbf{e}') = 0,$$

where  $\text{Var}()$  is a variance-covariance matrix, and  $\text{Cov}()$  is a covariance matrix.

Further,  $\mathbf{G}$  can be partitioned into other matrices for each factor contained within  $\mathbf{u}$ . Because  $\mathbf{u}$  will contain animal additive genetic effects,  $\mathbf{a}$ , and contemporary group effects,  $\mathbf{c}$ , then both  $\mathbf{a}$  and  $\mathbf{c}$  are assumed to follow normal distributions with null means. The variance-covariance matrix of  $\mathbf{a}$  is

$$\text{Var}(\mathbf{a}) = \mathbf{A}\sigma_a^2,$$

where  $\mathbf{A}$  is the numerator additive genetic relationship matrix of Wright (2), and  $\sigma_a^2$  is the additive genetic variance of the trait of interest. More will be given on the additive relationship matrix shortly.

Similarly,

$$\text{Var}(\mathbf{c}) = \mathbf{I}\sigma_c^2,$$

where  $\sigma_c^2$  is the contemporary group variance (variance of all contemporary group effects), and  $\mathbf{I}$  is an identity

matrix, which means the contemporary group effects are independent of each other. Finally,

$$\text{Cov}(\mathbf{a}, \mathbf{c}') = 0,$$

additive genetic and contemporary group effects are independent.

The residual effects are assumed to be independent of all other random factors, meaning their covariances with other random factors are 0. If known, the ratios of residual variance to the additive genetic variance and to the contemporary group variance should be provided, or at least the values that are intended to be used.

### Assumptions and Limitations

Limitations occur because of a lack of information in the data records. A factor that could be important is not included in the model because there is no information about that factor in the data files on animals. For example, an animal's record may or may not be affected by its health status. If the record was affected, then health status should be a factor in the linear model. However, health information may not be a part of the data files, and therefore, it cannot be included in the model. Consequently, an assumption is needed that animals were healthy when observations were taken. Depending on the trait and species, this assumption may or may not be critical.

A complete list of the explicit or implied assumptions should be a part of every model description, but often they are omitted. Readers of a scientific report may not be as familiar with the data files and production system as the authors, and may not be able to assess the assumptions that need to be made. The presence of this part of the model helps readers to judge the quality of an analysis.

### Genetic Relationships

The animal model works best when pedigree information is complete and accurate. Preparing pedigree files can be tedious, depending on the species. Animals can be registered in different organizations with different identifications in each organization, both of which may be found in the data files. Efforts are needed to make sure that each animal has only one unique identification in both the data files and the pedigree file.

Good identification systems include codes for breed of the animal and year of birth.

Next, pedigrees of animals need to be arranged in chronological order. Parents should appear in a list before (ahead of) their progeny. Ordering a pedigree is most easily accomplished by sorting animals by birthdate. Birthdates, however, can be incorrectly recorded, or for many individuals may not be available. One approach is to assume that all birthdates are incorrect. Assign all animals a generation number of 1. Then cycle through the pedigree file and modify the generation numbers so that the generation numbers of the sire and dam of an animal are at least one unit greater than the generation number of that animal. Iterate through the pedigree file as many times as needed until no further modifications are made to any generation numbers.

### Inbreeding Coefficients

Genetic relationships among individuals were worked out by Wright [6] as correlation coefficients. Later, only the numerators of these correlation coefficients were needed in genetic evaluation, and were called Wright's numerator, additive genetic relationships. The dimensions of the additive genetic relationship matrix,  $\mathbf{A}$ , equal the number of animals ( $N$ ) in the pedigree. The pedigree file usually contains more individuals than are represented with records in the data file. If constructed, the  $\mathbf{A}$  matrix would be 95% full (nonzero numbers). For one million animals, the storage of these numbers would be an overwhelming problem computationally. Fortunately, during genetic evaluation, only the inverse of this matrix is needed. The inverse is typically very sparse by comparison, and the elements in the matrix can be generated as needed and do not ever need to be stored. For this to work, the inbreeding coefficients of every animal need to be determined.

Henderson [7] showed that the additive relationship matrix could be written as

$$\mathbf{A} = \mathbf{T}\mathbf{B}\mathbf{T}',$$

where  $\mathbf{T}$  is a lower triangular matrix and  $\mathbf{B}$  is a diagonal matrix. The diagonals of  $\mathbf{B}$ , say  $b_{ii}$ , were shown to equal

$$b_{ii} = (0.5 - 0.25(F_s + F_d)),$$

where  $F_s$  and  $F_d$  are the inbreeding coefficients of the sire and dam, respectively, of the  $i$ th individual. If one parent is unknown, then

$$b_{ii} = (0.75 - 0.25F_p),$$

where  $F_p$  is the inbreeding coefficient of the parent that is known. Lastly, if neither parent is known, then  $b_{ii} = 1$ . An inbreeding coefficient indicates the proportion of alleles that are in common within an individual due to being inherited from common ancestors some generations back in the pedigree. This happens when related animals are mated together.

The key discovery of Henderson [7] was that

$$\mathbf{A}^{-1} = \mathbf{T}'^{-1}\mathbf{B}^{-1}\mathbf{T}^{-1},$$

and that each row of  $\mathbf{T}^{-1}$  had a one on the diagonal and two negative one-halves on the off-diagonals corresponding to the locations of the sire and dam of that individual. Only two nonzero numbers that are always equal to 1 or  $-\frac{1}{2}$ . The elements of  $\mathbf{B}^{-1}$  are equal to  $b_{ii}^{-1}$ . Every animal could have a different  $b_{ii}$ , so these would need to be stored.

Meuwissen and Luo [8] developed a very efficient algorithm to compute inbreeding coefficients, and from these come  $b_{ii}$ . The algorithm requires animals to be chronologically ordered and processed, so that inbreeding coefficients of an animal's ancestors are known before that animal is processed. Once the inbreeding coefficients are known, then  $b_{ii}$  are easily obtained. Having  $b_{ii}$  for each individual, then elements of  $\mathbf{A}^{-1}$  can be calculated readily, as needed. Consult Meuwissen and Luo [8] for details of their algorithm.

### Missing Parents

Animals with unknown parents are assumed to be animals from a large randomly mating population of unrelated, unselected individuals. This group of animals is known as the *base population*. In dairy cattle, the base population might be animals that were born in the 1950s. However, even today, there are animals in the data files and pedigree files that have unknown parentage. Clearly, these animals are genetically different from animals born 60 years earlier, and they should belong to a population different from the base population. For this reason, unknown parents are assigned to genetic

groups based on the year of birth of their progeny, and whether the progeny was male or female [9, 10]. Suppose an animal was born in 2009 and was a male. If the male parent of this animal was unknown, then it would be assigned to a Sire of Males group for 2009, and if the female parent was unknown assignment would be to a Dam of Males group for 2009. The assumption made for assignments is that the selection intensity of each pathway differs so that the genetic means of these groups would be different. The other two pathways are Sires of Females and Dams of Females. These groups may be further subdivided, based on breeds or countries of origin.

Computationally, genetic groups are treated as though they were a separate individual, which requires simple modifications in the computation of elements of  $\mathbf{A}^{-1}$ . Genetic groups are essential in the animal model in order to obtain unbiased estimates of genetic trends, and accurate evaluations of all animals. There are always animals with unknown parents resulting from movement of animals between countries, but also between herds or flocks within a country.

### Genetic Evaluation

Best linear unbiased prediction (or BLUP) has been used since 1970 when Henderson applied the method to genetic evaluation of dairy bulls in the northeastern United States even though the theory had been available since 1950. BLUP requires setting up and solving a linear system of equations that is equal in size to the number of animals and number of levels of other effects that are in the model. The equations are known as the mixed model equations (MME). The solutions to these equations are usually obtained by iteration rather than through the direct inverse of the large coefficient matrix. Many various strategies have been devised to iterate solutions quickly.

Iteration is a technique where initial solutions begin at zero, and by going through one equation at a time, each solution is updated based on the values of the other solutions at that time. An updated solution causes all subsequent solutions to change during the updating process. Iterations continue until the changes are less than a certain value (like  $1 \times 10^{-9}$ ), at which point the solutions have reached convergence.



## Mixed Model Equations

The mixed model equations (MME) of Henderson [2] yield the BLUP predictors of the random effects of the linear model and generalized least squares estimators of the fixed effects. Following the notation for the linear model, the MME, generally, are written as

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

From one model to the next,  $\mathbf{X}$ ,  $\mathbf{Z}$ ,  $\mathbf{R}$ , and  $\mathbf{G}$  change in how they may be constructed, but the MME always have the same form. Computations also differ and may simplify in certain cases.

## Estimation of Variances

Two methods of estimation of  $\mathbf{G}$  and  $\mathbf{R}$  are commonly used in animal breeding. First is Restricted (or Residual) Maximum Likelihood (REML) [11–13], of which there are four computational versions, and second is Bayesian Estimation using Gibbs Sampling as a tool to maximize the joint posterior distribution [14]. Both methods require major computing time, and consequently, smaller subsamples of the complete data file are chosen for estimating covariance matrices.

The main advantage of these two methods is that the estimated covariance matrices are positive definite matrices that can be used directly in MME. A covariance matrix must be positive definite like a variance must always be positive.

Another advantage of the methods is that the REML and Bayesian estimators are more accurate than other methods. The disadvantage is the increased computing demands necessitating the need to subsample the data into smaller sets.

## Reliabilities

All estimates of breeding values of animals require information about the accuracy or reliability of that estimate. Theoretically, standard errors of prediction are derived from the inverse of the coefficient matrix of the MME. Given that solutions result from iteration techniques rather than inversion, the standard errors of prediction must be approximated. Standard errors of prediction are often converted into a percent reliability

that goes from 0 to 100. There are many approximation methods in use.

Reliabilities depend on the number of progeny, the number of contemporaries of those progeny, the completeness of pedigree information, and the variance and covariance parameters. Depending on the reliability of an animal's Estimated Breeding Value, (EBV), the EBV may or may not be made officially public. Minimum standards are agreed upon by industry committees.

## Genetic Trends

Genetic trends are estimated by averaging animal EBVs by year of birth, or by years in which they make records. In dairy cattle, for example, the EBVs of all cows that were born in a given year and which completed at least one lactation can be averaged. By plotting these averages by year of birth, the genetic trend in milk production can be quantified. Plotting the average EBVs of dairy bulls by their year of birth would give a different trend, reflecting how bulls were chosen to be in artificial insemination. The cow averages would reflect how those bulls were used in breeding programs, and would likely lag a couple of years behind the sire birth years. Graphs of genetic trends should have details of what the averages represent.

## Single Record per Animal

The models presented in this and the following sections become progressively more complicated, due to the type of traits that are considered. First, a simple animal model is illustrated which covers traits that are observed only once in an animal's lifetime, such as age at first breeding. Next, the situation where animals can be observed more than once per lifetime, such as annual antler production in elk, wool production in sheep, or race results in horses. Then traits influenced by maternal effects, such as birth weights, weaning weights, and calving ease, are considered. Longitudinal data, traits observed over time, such as milk production or egg production, provide models that analyze the shapes of curves. Multiple trait models are applied to two or more traits at a time, and include genetic and environmental correlations among the traits. Threshold models for categorical data are described, and finally, a model for the analysis of survival data is presented.

Data

The case of a single record per animal is presented in some detail because this is the simplest model and illustrates the process of constructing mixed model equations, and what happens to the solutions to those equations. The models are presented in a generic fashion without reference to particular species.

Table 1 contains the pedigree information and data on 16 animals. The first four animals are base population animals without records and without known parents. The inbreeding coefficients are shown and the values of  $b_{ii}$  which come from the inbreeding coefficients. Records were made over 2 years, and there were two contemporary groups within each year. Assume that the residual variance is 600, the genetic variance is 300, and the variance of contemporary group effects is 100. Heritability (denoted as  $h^2$ ) is defined as the

genetic variance divided by the total phenotypic variance, which is the sum of the genetic, contemporary group, and residual variances. In this case,  $h^2 = 0.3$ .

Mixed Model Equations

The equation of the model is (in scalar form)

$$y_{ijk} = (YR)_i + (CG)_j + a_k + e_{ijk},$$

where

$y_{ijk}$  are the observations on animal  $k$ , belonging to contemporary group,  $(CG)_j$ , born in year,  $(YR)_i$ ,

$$\mathbf{y} = \{y_{ijk}\}.$$

$(YR)_i$  are the fixed, year effects,

$$\mathbf{b} = \{(YR)_i\}.$$

Animal Breeding, Modeling in. Table 1 Example data for simple animal model

| Animal | Sire | Dam | $F_i$   | $b_{ii}$ | Year | CG | Record |
|--------|------|-----|---------|----------|------|----|--------|
| 1      | 0    | 0   | 0.00000 | 1.00000  |      |    |        |
| 2      | 0    | 0   | 0.00000 | 1.00000  |      |    |        |
| 3      | 0    | 0   | 0.00000 | 1.00000  |      |    |        |
| 4      | 0    | 0   | 0.00000 | 1.00000  |      |    |        |
| 5      | 1    | 2   | 0.00000 | 0.50000  | 1    | 1  | 78     |
| 6      | 3    | 4   | 0.00000 | 0.50000  | 1    | 1  | 26     |
| 7      | 3    | 2   | 0.00000 | 0.50000  | 1    | 1  | 111    |
| 8      | 1    | 4   | 0.00000 | 0.50000  | 1    | 1  | 122    |
| 9      | 5    | 6   | 0.00000 | 0.50000  | 1    | 1  | 98     |
| 10     | 7    | 8   | 0.00000 | 0.50000  | 1    | 2  | 48     |
| 11     | 5    | 8   | 0.12500 | 0.50000  | 1    | 2  | 109    |
| 12     | 7    | 6   | 0.12500 | 0.50000  | 1    | 2  | 94     |
| 13     | 1    | 6   | 0.00000 | 0.50000  | 1    | 2  | 103    |
| 14     | 3    | 8   | 0.00000 | 0.50000  | 2    | 3  | 78     |
| 15     | 5    | 4   | 0.00000 | 0.50000  | 2    | 3  | 69     |
| 16     | 7    | 8   | 0.00000 | 0.50000  | 2    | 3  | 44     |
| 17     | 9    | 6   | 0.25000 | 0.50000  | 2    | 3  | 12     |
| 18     | 11   | 10  | 0.18750 | 0.46875  | 2    | 4  | 54     |
| 19     | 11   | 12  | 0.06250 | 0.43750  | 2    | 4  | 89     |
| 20     | 9    | 8   | 0.12500 | 0.50000  | 2    | 4  | 82     |

$(CG)_j$  are the random, contemporary group effects,

$$\mathbf{c} = \{(CG)_j\}.$$

$a_k$  are the random, additive genetic effects of individual animals,

$$\mathbf{a} = \{a_k\}$$

$e_{ijk}$  are random, residual effects.

Also,

$$\mathbf{u} = \begin{pmatrix} \mathbf{c} \\ \mathbf{a} \end{pmatrix},$$

and

$$\text{Var} \begin{pmatrix} \mathbf{c} \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_4 \sigma_c^2 & 0 \\ 0 & \mathbf{A} \sigma_a^2 \end{pmatrix} = \mathbf{G}.$$

Finally,

$$\mathbf{R} = \mathbf{I}_{16} \sigma_e^2,$$

and  $\mathbf{Z}$  can be partitioned into one matrix for contemporary groups, and one for animal additive genetic values, i.e.,

$$\mathbf{Z} = (\mathbf{Z}_{cg} \quad \mathbf{Z}_a).$$

### Inverse of Relationship Matrix

Notice that the MME contain  $\mathbf{G}^{-1}$ , which equals

$$\mathbf{G}^{-1} = \begin{pmatrix} \mathbf{I}_{\frac{1}{\sigma_g^2}} & 0 \\ 0 & \mathbf{A}^{-1} \frac{1}{\sigma_a^2} \end{pmatrix}.$$

Thus, the inverse of the relationship matrix is needed. This matrix can be constructed readily following simple rules that Henderson [7] provided. For each animal with both parents known, 9 numbers are added into the inverse matrix. For an animal with only one parent known, 4 numbers are added, and for an animal with both parents unknown, 1 number is added. Start with a matrix of order 20 that is completely null. Then process the pedigree file, one animal at a time until all animals are included.

**Step 1:** For animal  $i$ , let  $\delta = b_{ii}^{-1}$ . For animal 1,  $\delta = 1$ ; for animal 6,  $\delta = 2$ , and for animal 19,  $\delta = 2.2857$ .

**Step 2:** Add  $\delta$  to the diagonal element for that animal. For animal 19, add  $\delta$  to element (19,19) of  $\mathbf{A}^{-1}$ .

**Step 3:** If the male parent is known, subtract  $-0.5\delta$  from elements  $(i, s)$  and  $(s, i)$ , where  $i$  is the animal's number and  $s$  is the sire's number. For example, for animal 19,  $i = 19$  and  $s = 11$ . Also add  $0.25\delta$  to element  $(s, s)$ .

**Step 4:** If the female parent (denoted by  $d$ ) is known, subtract  $-0.5\delta$  from elements  $(i, d)$  and  $(d, i)$ , and add  $0.25\delta$  to element  $(d, d)$ .

**Step 5:** If both parents are known add  $0.25\delta$  to elements  $(s, d)$  and  $(d, s)$ .

Numbers are accumulative as the pedigrees are processed. Many of the elements will stay null values, especially for animals that do not have progeny.

### Solutions to Equations

The resulting mixed model equations have 26 rows and columns with 26 unknowns to be estimated (2 year effects, 4 contemporary groups, and 20 animal effects). Even for this small example, to display the equations fully would take up too much space. The solutions to the MME were as follows:

|        |   |        |
|--------|---|--------|
| Year 1 | = | 88.8   |
| Year 2 | = | 61.6   |
| CG 1   | = | .015   |
| CG 2   | = | -.015  |
| CG 3   | = | -3.303 |
| CG 4   | = | 3.303  |

| Animal | EBV    | Reliability |
|--------|--------|-------------|
| 1 =    | 8.32   | .13         |
| 2 =    | 1.06   | .11         |
| 3 =    | -5.68  | .14         |
| 4 =    | -3.69  | .10         |
| 5 =    | 4.38   | .23         |
| 6 =    | -15.50 | .25         |
| 7 =    | -.95   | .25         |
| 8 =    | 7.37   | .23         |
| 9 =    | -4.28  | .20         |
| 10 =   | -6.58  | .18         |
| 11 =   | 9.49   | .24         |
| 12 =   | -3.79  | .24         |
| 13 =   | -.02   | .17         |
| 14 =   | 4.61   | .17         |
| 15 =   | 2.41   | .18         |
| 16 =   | -.30   | .16         |
| 17 =   | -17.17 | .25         |
| 18 =   | -.90   | .21         |
| 19 =   | 6.66   | .16         |
| 20 =   | 4.65   | .18         |

The solutions for the animal additive genetic effects are known as *Estimated Breeding Values* or EBVs.

Progeny are expected to inherit an average of the EBVs of its parents, on average. EBVs are sorted from highest to lowest, or best animal to poorest animal. Every animal in the pedigree file obtains a solution due to genetic relationships to progeny and other individuals. The EBV of an animal consists of combined information from the animal's parents, its progeny, and its own performance record. The BLUP methodology combines the information in an optimal manner to maximize the correlation of the EBV with the animal's true genetic merit. The reliability is obtained from the inverse of the MME coefficient matrix, and expressed as a correlation coefficient. The greater is the reliability, the more certainty in the ranking of the animals. The highest reliability was 0.25, for animals 6, 7, and 17. Note that animal 6 had 4 progeny, plus a record on itself, and both parents were known. Animal 7 had 3 progeny. Animal 17 had no progeny, but was an offspring of animal 6 and was inbred the most. Reliability also reflects the number of contemporaries that each animal has, but the differences in this example were not great. The reliabilities of all animals in this example are very small, but there were only 16 observations in total. Dairy bulls, for example, can have hundreds or thousands of progeny giving them reliabilities above 0.99. A minimum reliability level is chosen before EBVs are released to the public.

The solutions for the year effects were 88.8 and 61.6, respectively. Thus, performances were lower in year 2 compared to year 1. There is likely a reason for this difference, such as year 2 being hotter, or a shortage of good feed, or feed prices may have caused many animals to be removed from farms. The difference between years is not genetic.

Contemporary group effects were random in the model, and the solutions have an average of zero as a consequence of how they were included in the model. The CG solutions average zero within the year effects, due to the fact they were a factor nested within year effects.

Genetic trends can be computed from the EBVs. Because animals have only one record each, the genetic averages for years 1 and 2 are the average EBVs of animals 5–13 for year 1, and animals 14–20 for year 2. These give  $-1.10$  for year 1 and  $-0.01$  for year 2. Thus, the animals in year 2 were slightly better than in year 1, genetically.

## Repeated Records Animal Model

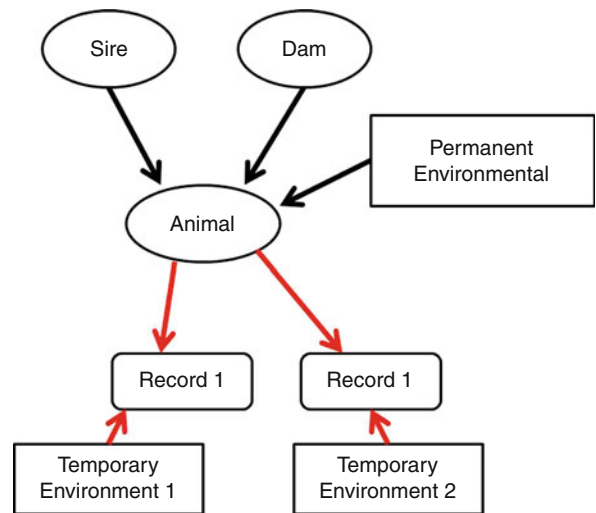
Animals are often observed more than once for some traits, such as

- Fleece weight of sheep in different years
- Calf records of a beef cow over time
- Test day records within a lactation for a dairy cow
- Litter size of sows over time
- Antler size of deer in different seasons
- Racing results of horses from several races

Animals are influenced by their environments, such as an athlete changes due to training and practice. These effects are called permanent environmental effects, and they accompany the animal every time the animal is observed for that trait as in Fig. 2. Permanent environmental effects are not genetic, in the sense that the animal does not transmit these effects to any of its offspring. The following figure illustrates a model to describe permanent environmental effects. Only with repeated records can the permanent environmental and genetic effects be separated and estimated.

## Data

In the example data of Table 2, animals have either one, two, or three records made in different years.



**Animal Breeding, Modeling in. Figure 2**

Diagram for an animal with repeated records including a permanent environmental effect

**Animal Breeding, Modeling in. Table 2** Example data for repeated records animal model

| Animal | Sire | Dam | Age  | Year 1 | Year 2 | Year 3 |
|--------|------|-----|------|--------|--------|--------|
|        |      |     | Yr 1 |        |        |        |
| 1      |      |     | 3    | 22     | 35     |        |
| 2      |      |     | 3    | 31     | 46     |        |
| 3      |      |     | 3    | 44     | 24     |        |
| 4      |      |     | 3    | 53     |        |        |
| 5      |      |     | 2    | 61     | 57     | 42     |
| 6      |      |     | 2    | 32     |        | 39     |
| 7      | 1    | 2   | 2    | 39     | 51     | 62     |
| 8      | 3    | 4   | 1    | 48     | 72     |        |
| 9      | 5    | 6   | 1    | 71     |        | 96     |
| 10     | 1    | 4   | 1    | 37     | 56     | 47     |
| 11     | 3    | 6   | 1    | 66     |        | 86     |
| 12     | 1    | 2   | 1    |        | 46     | 38     |

The contemporaries for each year differ, but are confounded with the year effect in this example. If the animals could be assigned to different herds or management groups, then contemporary groups and years would not be confounded.

Normally performance of animals either improves or declines with the age of the animal, so that ages of animals should be known. The age of the animal in year 1 is given in the table, and so ages range from 1 to 4 years.

### Mixed Model Equations

The equation of the model is (in scalar form)

$$y_{ijkl} = (YR)_i + (Age)_j + a_k + p_k + e_{ijkl},$$

where

$y_{ijkl}$  is observation  $l$  on animal  $k$ , belonging to age group  $j$ , born in year  $i$ .

$(YR)_i$  are fixed, year effects.

$(Age)_j$  are fixed, age group effects.

$a_k$  are random, additive genetic effects of individual animals.

$p_k$  are random, permanent environmental effects of individual animals.

$e_{ijk}$  are random, residual effects.

The covariance matrices of the random variables (animals, permanent environmental, and residual effects, respectively) are

$$\text{Var} \begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma_a^2 & 0 & 0 \\ 0 & \mathbf{I}\sigma_p^2 & 0 \\ 0 & 0 & \mathbf{I}\sigma_e^2 \end{pmatrix}.$$

The total variance is

$$\sigma_y^2 = \sigma_a^2 + \sigma_p^2 + \sigma_e^2,$$

and the heritability is

$$h^2 = \frac{\sigma_a^2}{\sigma_y^2}.$$

Repeatability is a measure of the average similarity of repeated records on animals across the population (part genetic and part environmental), and is defined as a ratio of variances as

$$r = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_y^2},$$

which is always going to be greater than or equal to heritability.

Let

$\mathbf{Xb}$  represent the fixed factors of years and ages

$\mathbf{Za}$  represent the random animal additive genetic effects

$\mathbf{Wp}$  represent the random animal permanent environmental effects

then the mixed model equations may be written as

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k_a & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W} + \mathbf{I}k_p \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{pmatrix},$$

and let

$$k_a = \sigma_e^2 / \sigma_a^2 = 1.33333, \quad \text{and} \quad k_p = \sigma_e^2 / \sigma_p^2 = 3.$$

There are 26 observations in the example of Table 2, with 3 years and 4 age groups represented, plus 12 animal additive genetic effects and 12 permanent environmental effects giving a total of 31 equations.



The **A** matrix is simple to construct because none of the animals are inbred. The resulting solutions are given in Table 3.

There are two possible uses of these solutions. First is to rank the animals for their genetic ability in order to plan future matings using the Estimated Breeding Values. Secondly, there is the decision about which animals to keep to make another record. The criterion for making this decision is the Most Probable Producing Ability, [1] which is the sum of the genetic and permanent environmental solutions (shown in the last column of Table 3). Thus, Animal 9 would likely make the best future performance of those 12 animals and would also likely generate the best future progeny.

Comments

One could ask if permanent environmental effects are permanent. The answer is yes, but as an animal ages, it encounters new permanent environmental effects which accumulate with the previous effects. Hence, permanent environmental effects are cumulative over the life of an animal. This means permanent environmental effects are not constant throughout an animal's life. With the model as described in this section, the

Animal Breeding, Modeling in. Table 3 A set of solutions to the example data on repeated records model

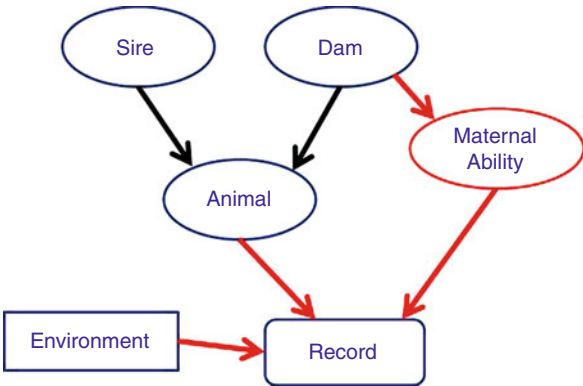
| Years |       | Ages |       | Animal  |            | MPPA  |
|-------|-------|------|-------|---------|------------|-------|
|       |       |      |       | Genetic | Perm. Env. |       |
| 1     | 0.00  | 1    | 50.92 | 1       | −8.37      | 8.64  |
| 2     | 18.34 | 2    | 46.01 | 2       | −1.50      | −0.52 |
| 3     | 29.08 | 3    | 37.24 | 3       | 1.06       | −1.84 |
|       |       | 4    | 19.50 | 4       | 3.14       | 3.16  |
|       |       |      |       | 5       | 4.96       | −0.84 |
|       |       |      |       | 6       | 0.71       | −5.00 |
|       |       |      |       | 7       | −2.94      | 1.77  |
|       |       |      |       | 8       | 2.18       | 0.07  |
|       |       |      |       | 9       | 9.68       | 6.08  |
|       |       |      |       | 10      | −6.66      | −3.60 |
|       |       |      |       | 11      | 6.00       | 4.55  |
|       |       |      |       | 12      | −10.64     | −5.07 |

assumption is that permanent environmental effects are constant. Some of the cumulative parts, therefore, flow into the temporary environmental effects, and some are averaged with the previous permanent environmental effects.

Another assumption is that the genetic component of each record on an animal is the same. Genes are known to change in activity as an animal ages due to age, but also due to epigenetics (environmental effects that cause change to an animal's DNA). Thus, repeated records on one animal could have different genetic components. That means the genetic correlation between records is less than unity. A better model would be to assume that all records are genetically different (but correlated) traits. This would take into account both the different genetic effects associated with each record, and also the accumulation of permanent environmental effects with time.

Maternal Effects Animal Model

In mammalian species of livestock, such as beef cattle, sheep, or swine, the female provides an environment for its offspring to survive and grow in terms of protection and nourishment. Figure 3 illustrates how maternal effects can affect offspring records. Females vary in their ability to provide a suitable environment for their offspring, and this variability has a genetic basis. Offspring directly inherit an ability to grow (or survive) from both parents, and environmentally do better or poorer depending on their dam's genetic



Animal Breeding, Modeling in. Figure 3 Diagram illustrating maternal genetic effects

maternal ability. Maternal ability is a genetic trait expressed by the dam in the offsprings' performance, and is transmitted, like all genetic traits, from both parents. Maternal ability is only expressed by females when they have offspring (i.e., much like milk yield in dairy cows) [15, 16].

### Data

The example data of Table 4 are weights on animals at an early age.

A model to account for maternal ability is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{p} + \mathbf{e},$$

where  $\mathbf{y}$  is the growth trait of a young animal,  $\mathbf{b}$  is a vector of fixed factors influencing growth, in this case contemporary group effects,  $\mathbf{a}$  is a vector of random animal additive genetic effects (i.e., direct genetic effects),  $\mathbf{m}$  is a vector of random maternal genetic (dam) effects, and  $\mathbf{p}$ , in this model, is a vector of maternal permanent environmental effects (because dams may have more than one offspring in the data – repeated records).

The expectations of the random vectors,  $\mathbf{a}$ ,  $\mathbf{m}$ ,  $\mathbf{p}$ , and  $\mathbf{e}$  are all null vectors in a model without selection, and the variance-covariance structure is

$$\text{Var} \begin{pmatrix} \mathbf{a} \\ \mathbf{m} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{A}\sigma_{am} & 0 & 0 \\ \mathbf{A}\sigma_{am} & \mathbf{A}\sigma_m^2 & 0 & 0 \\ 0 & 0 & \mathbf{I}\sigma_p^2 & 0 \\ 0 & 0 & 0 & \mathbf{I}\sigma_e^2 \end{pmatrix},$$

where  $\sigma_a^2$  is the additive genetic variance,  $\sigma_m^2$  is the maternal genetic variance,  $\sigma_{am}$  is the covariance

between additive and maternal genetic effects, and  $\sigma_p^2$  is the maternal permanent environmental variance. Also,

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{m} \end{pmatrix} \mid \mathbf{A}, \mathbf{G} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{G} \otimes \mathbf{A} \right),$$

where

$$\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix},$$

and

$$\mathbf{p} \mid \mathbf{I}, \sigma_p^2 \sim N(0, \mathbf{I}\sigma_p^2),$$

and

$$\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2).$$

In this model, a female animal,  $i$ , could have its own growth record for estimating  $\hat{a}_i$ . The same female could later have offspring for estimating  $\hat{m}_i$  and  $\hat{p}_i$ , and the offspring would also contribute toward  $\hat{a}_i$ . The maternal effects model can be more complicated if, for example, embryo transfer is practiced. Recipient dams would have maternal effects, but would not have direct genetic effects on that calf [17].

### Mixed Model Equations

The MME are represented as

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 \\ \mathbf{Z}'_1\mathbf{X} & \mathbf{Z}'_1\mathbf{Z}_1 + \mathbf{A}^{-1}k_{11} & \mathbf{Z}'_1\mathbf{Z}_2 + \mathbf{A}^{-1}k_{12} \\ \mathbf{Z}'_2\mathbf{X} & \mathbf{Z}'_2\mathbf{Z}_1 + \mathbf{A}^{-1}k_{12} & \mathbf{Z}'_2\mathbf{Z}_2 + \mathbf{A}^{-1}k_{22} \\ \mathbf{Z}'_3\mathbf{X} & \mathbf{Z}'_3\mathbf{Z}_1 & \mathbf{Z}'_3\mathbf{Z}_2 \\ \mathbf{X}'\mathbf{Z}_3 & \mathbf{Z}'_1\mathbf{Z}_3 & \mathbf{Z}'_2\mathbf{Z}_3 \\ \mathbf{Z}'_3\mathbf{Z}_3 + \mathbf{I}k_{33} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{m}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'_1\mathbf{y} \\ \mathbf{Z}'_2\mathbf{y} \\ \mathbf{Z}'_3\mathbf{y} \end{pmatrix},$$

where

$$\begin{aligned} \begin{pmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{pmatrix} &= \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix}^{-1} \sigma_e^2, \\ &= \begin{pmatrix} 49 & -7 \\ -7 & 26 \end{pmatrix}^{-1} (81), \\ &= \begin{pmatrix} 1.7192 & 0.4628 \\ 0.4628 & 3.2400 \end{pmatrix}. \end{aligned}$$

**Animal Breeding, Modeling in. Table 4** Example data for maternal genetic effects model

| Animal | Sire | Dam | CG | Weight |
|--------|------|-----|----|--------|
| 5      | 1    | 3   | 1  | 156    |
| 6      | 2    | 3   | 1  | 124    |
| 7      | 1    | 4   | 1  | 135    |
| 8      | 2    | 4   | 2  | 163    |
| 9      | 1    | 3   | 2  | 149    |
| 10     | 2    | 4   | 2  | 138    |

The solutions to the MME are

$$\hat{\mathbf{b}} = \begin{pmatrix} 137.8469 \\ 150.4864 \end{pmatrix}, \quad \hat{\mathbf{p}} = \begin{pmatrix} 0.0658 \\ -0.0658 \end{pmatrix},$$

$$\hat{\mathbf{a}} = \begin{pmatrix} 2.3295 \\ -2.3295 \\ 0.1280 \\ -0.1280 \\ 5.1055 \\ -4.1143 \\ 0.2375 \\ 2.0161 \\ 0.5447 \\ -3.7896 \end{pmatrix}, \quad \text{and} \quad \hat{\mathbf{m}} = \begin{pmatrix} -0.3328 \\ 0.3328 \\ 0.1646 \\ -0.1646 \\ -0.6379 \\ 0.6792 \\ -0.1254 \\ -0.3795 \\ 0.0136 \\ 0.4499 \end{pmatrix}.$$

### Comments

Maternal genetic models require a good data structure to be successful [18]. That means that there should be many females with weight records in the data who also have several progeny with weight records in the data. In this way direct genetic and maternal genetic effects can be efficiently separated during estimation. In the example data, the structure is not good because the females that were dams did not have any weight records on themselves. Hence the strong negative relationship between direct and maternal genetic estimates, which means the maternal genetic estimates are based mostly on the direct genetic estimates and the prior genetic correlation that was assumed, which was negative.

Sire structure is also important in that sires should have many daughters that have also had their own progeny. The maternal genetic ability of a sire's daughters cannot be accurately estimated without those daughters displaying their maternal ability on their own progeny. In many studies or application of maternal genetic effects models, the data structure is too poor from which to estimate variances and covariances of direct and maternal genetic effects.

### Random Regression Animal Model

All biological creatures grow and perform over their lifetime. Traits that are measured at various times during that life are known as *longitudinal* data. Examples are body weights [19], body lengths, milk production [20], feed intake, fat deposition, and egg production

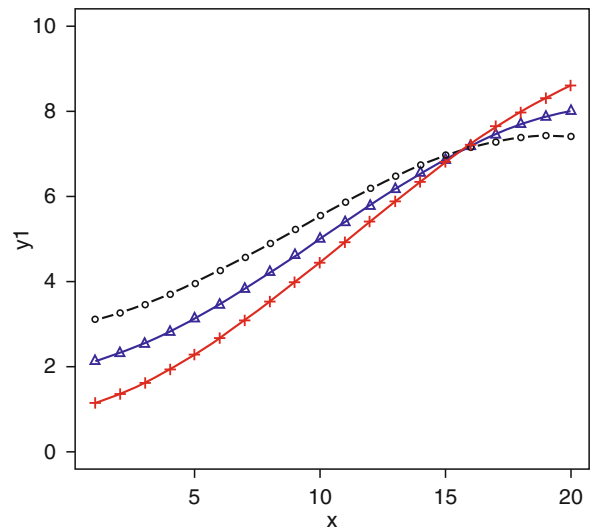
[21]. On a biological basis, there could be different genes that turn on or turn off as an animal ages causing changes in physiology and performance. The time variable (or age) can be recorded in years, months, weeks, days, hours, minutes, or seconds, so that, in effect, there could be a continuum or continuous range of points in time when an animal could be observed for a trait. These traits have also been called *infinitely dimensional* traits.

If observations were plotted on a graph where the  $x$ -axis is time and the  $y$ -axis is the magnitude of the observations, then a *trajectory* is obtained for a group of animals. However, not every individual will follow the average trajectory, see Fig. 4. There are variations in the shape of their trajectory, and the breeder may wish to change the shape for animals in his(her) production system in order to be more profitable.

The average trajectory, in black, is given by the equation

$$f(t_i) = A + Bt_i + Ct_i^2 + Dt_i^3,$$

where  $A = 2$ ,  $B = 0.1$ ,  $C = 0.03$ , and  $D = -0.001$ . The red and blue trajectories are extremes for animals in the population. The assumption is that  $f(t_i)$  is the average trajectory and every animal will have its own trajectory,



Animal Breeding, Modeling in. Figure 4

Trajectories of different animals for a trait measured over time

that means every animal will have different  $A$ ,  $B$ ,  $C$ , and  $D$  values. Because these parameters are regression coefficients, the model becomes the *random regression* model, and every animal will have four regression coefficients to be estimated, as deviations from the average trajectory regression coefficients.

### Regression Functions

A problem in using time covariates to various powers is that the numbers can become very large, very quickly. For example, if the time variable,  $t$ , goes from 1 to 20, then  $t^3$  would range from 1 to 8,000. In least squares like equations, the diagonal element for that variable would be  $(8,000)^2$ . Then there could be many thousands of observations. The large numbers can lead to serious rounding errors and may cause problems in solving the equations.

Another problem is that there will be high correlations among the time variables in the function, because they are all based on the same  $t$  value. This may also lead to near singularity and to problems in solving the equations. A solution for the above problems is to use Legendre polynomials which convert the time variable,  $t$ , into covariates that are firstly scaled to be between  $-1$  and  $+1$ , and then converted to be independent of each other.

**Scaling Time Variables** Time variables have to be standardized to the interval between  $-1$  and  $+1$ . The formula to standardize  $t_\ell$  is

$$x_\ell = -1 + 2 \left( \frac{t_\ell - t_{\min}}{t_{\max} - t_{\min}} \right).$$

**Legendre Polynomials** The first two Legendre polynomials are defined as

$$P_0(x) = 1, \quad \text{and}$$

$$P_1(x) = x,$$

then, in general, the  $n + 1$  polynomial is described by the following recursive equation:

$$P_{n+1}(x) = \frac{1}{n+1} ((2n+1)xP_n(x) - nP_{n-1}(x)).$$

These quantities are “normalized” using

$$\phi_n(x) = \left( \frac{2n+1}{2} \right)^{0.5} P_n(x).$$

This gives the following series:

$$\phi_0(x) = \left( \frac{1}{2} \right)^{0.5} P_0(x) = 0.7071$$

$$\begin{aligned} \phi_1(x) &= \left( \frac{3}{2} \right)^{0.5} P_1(x) \\ &= 1.2247x \end{aligned}$$

$$P_2(x) = \frac{1}{2} (3xP_1(x) - 1P_0(x))$$

$$\begin{aligned} \phi_2(x) &= \left( \frac{5}{2} \right)^{0.5} \left( \frac{3}{2}x^2 - \frac{1}{2} \right) \\ &= -0.7906 + 2.3717x^2, \end{aligned}$$

and so on. The first six can be put into a matrix,  $\Lambda$ , as

$$\Lambda' = \begin{pmatrix} 0.7071 & 0 & 0 \\ 0 & 1.2247 & 0 \\ -0.7906 & 0 & 2.3717 \\ 0 & -2.8062 & 0 \\ 0.7955 & 0 & -7.9550 \\ 0 & 4.3973 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 4.6771 & 0 & 0 \\ 0 & 9.2808 & 0 \\ -20.5206 & 0 & 18.4685 \end{pmatrix}.$$

Now define a vector,  $\mathbf{m}$ , containing the polynomials of standardized time values,

$$\mathbf{m}' = (1 \ x \ x^2 \ x^3 \ x^4 \ x^5).$$

The covariates to use in the model are equal to

$$\Lambda' \mathbf{m}.$$

To illustrate, suppose time goes from 10 to 60 days over which animals are observed for their growth,  $t_{\min} = 10$  and  $t_{\max} = 60$ . An animal is observed on day 43. The standardized time variable is

$$x = -1 + 2 \frac{(43 - 10)}{(60 - 10)} = -0.32,$$

and

$$\mathbf{m}' = (1 \ -0.32 \ 0.1024 \ -0.0328 \ 0.01048 \ -0.00336).$$

Finally,

$$\Lambda' \mathbf{m} = \begin{pmatrix} 0.7071 \\ -0.3919 \\ -0.5477 \\ 0.7446 \\ 0.0782 \\ -0.7961 \end{pmatrix}.$$

The order of Legendre polynomials is equal to the highest power of  $x$ . Research is needed to determine the best order of fit for any given situation.

### Random Regression Model

Like other animal models, the random regression model (RRM) accounts for years, contemporary groups, and animal additive genetic effects, but in addition, it needs to account for the following:

- Curves for different groups of animals, such as age groups, month of calving groups, and breeds which could have different shapes of curves. These factors would be fixed effects in the model. The curves may be fit using Legendre polynomials, or other functions of time, or as classification variables with many levels.
- Curves for each individual animal, using Legendre polynomials, which are random factors in the model. Each animal has a number of regression coefficients to be estimated. For each animal, there are additive genetic parameters as well as permanent environmental parameters.
- The possibility of residual variances changing over the observable time period. For example, as animals grow, their mean weight increases and so does the variance of weights at a given age.

A simplified RRM for a single trait can be written as

$$y_{ijkn:t} = F_i + g(t)_j + r(a, x, m1)_k + r(pe, x, m2)_k + e_{ijkn:t},$$

where

$y_{ijkn:t}$  is the  $n$ th observation on the  $k$ th animal at time  $t$  belonging to the  $i$ th fixed factor and the  $j$ th group.

$F_i$  is a fixed effect that is independent of the time scale for the observations, such as a cage effect, a location effect, or a herd-test date effect.

$g(t)_j$  is a function or functions that account for the phenotypic trajectory of the average observations across all animals belonging to the  $j$ th group.

$r(a, x, m1)_k = \sum_{\ell=0}^{m1} a_{k\ell} x_{ijk:\ell}$  is the notation adopted for a random regression function. In this case,  $a$  denotes the additive genetic effects of the  $k$ th animal,  $x$  is the vector of time covariates, and  $m1$  is the order of the regression function. So that  $x_{ijk:\ell}$  are the covariables related to time  $t$ , and  $a_{k\ell}$  are the animal additive genetic regression coefficients to be estimated.

$r(pe, x, m2)_k = \sum_{\ell=0}^{m2} p_{k\ell} x_{ijk:\ell}$  is a similar random regression function for the permanent environmental ( $pe$ ) effects of the  $k$ th animal.

$e_{ijkn:t}$  is a random residual effect with mean null and with possibly different variances for each  $t$  or functions of  $t$ .

The function,  $g(t)_j$ , can be either linear or nonlinear in  $t$ . Such a function is necessary in a RRM to account for the phenotypic relationship between  $y$  and the time covariables (or other types of covariables that could be used in a RRM). In a test day model,  $g(t)_j$  accounts for different lactation curve shapes for groups of animals defined by years of birth, parity number, and age and season of calving within parities, for example. With growth data,  $g(t)_j$  accounts for the growth curve of males or females of breed X or breed Y from young or old dams.

If the shape of the phenotypic relationship is not known or is nonlinear, then  $g(t)_j$  could be a set of classification variables. Classification variables take up more degrees of freedom and require a large number of observations per level, but they do not force the user to explicitly define the shape of the trajectory. A mathematical function, on the other hand, does not use many degrees of freedom and gives a smooth trajectory over time regardless of the number of observations. The choice of classification variables or mathematical function is up to the researcher. If data are very numerous, and the mathematical function fits the data well, then either approach will generally lead to the same results. The phenotypic relationships,  $g(t)_j$ ,



are important to a RRM analysis and deserve care and effort in their correct specification.

**Mixed Model Equations** In matrix notation, the RRM is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e},$$

where  $\mathbf{b}$  contains  $F_i$  and  $g(t)_j$  effects,  $\mathbf{a}$  contains  $m_1 + 1$  additive genetic regression coefficients for each animal,  $\mathbf{p}$  contains  $m_2 + 1$  permanent environmental regression coefficients for each animal with data, and  $\mathbf{e}$  contains the temporary environmental effects. Also,

$$\text{Var} \begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A} \otimes \mathbf{G} & 0 & 0 \\ 0 & \mathbf{I} \otimes \mathbf{P} & 0 \\ 0 & 0 & \mathbf{R} \end{pmatrix},$$

where  $\mathbf{G}$  is the variance-covariance matrix of the additive genetic random regression coefficients of order  $m_1 + 1$ ;  $\otimes$  if the direct product operator which multiplies every element of  $\mathbf{A}$  by the matrix  $\mathbf{G}$ ;  $\mathbf{P}$  is the variance-covariance matrix of the permanent environmental random regression coefficients of order  $m_2 + 1$ ; and  $\mathbf{R}$  is a diagonal matrix of temporary environmental variances which could vary depending on  $t$ , or  $\mathbf{R}$  could be block diagonal with an autocorrelation structure for each animal's records. The mixed model equations (MME) are represented as

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_1 \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{A}^{-1} \otimes \mathbf{G}^{-1} \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_1 \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_2 & \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_2 & \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{I} \otimes \mathbf{P}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

## Example Data Analysis by RRM

A very simplified example is given below to illustrate the degree of complexity of RRM. Four animals were observed multiple times at different ages for a trait, as shown in Table 5.

The model equation might be

$$\begin{aligned} y_{jik:t} = & V_j + b_0 + b_1(A) + b_2(A)^2 \\ & + (a_{i0}z_0 + a_{i1}z_1 + a_{i2}z_2) \\ & + (p_{i0}z_0 + p_{i1}z_1 + p_{i2}z_2) + e_{jik:t} \end{aligned}$$

where

$V_j$  is a random contemporary group (visit) effect which is assumed to follow a normal distribution with mean 0 and variance,  $\sigma_c^2 = 4$ .

$b_0$ ,  $b_1$ , and  $b_2$  are fixed regression coefficients on  $(A) = \text{age}$  and  $\text{age squared}$  which describes the general relationship between age and the observations.

$a_{i0}$ ,  $a_{i1}$ , and  $a_{i2}$  are random regression coefficients for animal  $i$  additive genetic effects, assumed to follow a multivariate normal distribution with mean vector null and variance-covariance matrix,  $\mathbf{G}$ .

$p_{i0}$ ,  $p_{i1}$ , and  $p_{i2}$  are random regression coefficients for animal  $i$  permanent environmental effects, assumed to follow a multivariate normal distribution with mean vector null and variance-covariance matrix,  $\mathbf{P}$ .

$z_0$ ,  $z_1$ , and  $z_2$  are the Legendre polynomials based on standardized ages and derived as indicated earlier. The minimum age was set at 18 and the maximum age was set at 68 for calculating the Legendre polynomials.

and  $e_{jik}$  is a temporary residual error term assumed to follow a normal distribution with mean 0 and variance,  $\sigma_e^2 = 9$ . In this example, the residual variance is assumed to be constant across ages.

**Animal Breeding, Modeling in.** Table 5 Example data for random regression model

| Cow | Sire | Dam | Visit 1 |      | Visit 2 |      | Visit 3 |      | Visit 4 |      |
|-----|------|-----|---------|------|---------|------|---------|------|---------|------|
|     |      |     | Age(m)  | Obs. | Age(m)  | Obs. | Age(m)  | Obs. | Age(m)  | Obs. |
| 1   | 7    | 5   | 22      | 224  | 34      | 236  | 47      | 239  |         |      |
| 2   | 7    | 6   | 30      | 244  | 42      | 247  | 55      | 241  | 66      | 244  |
| 3   | 8    | 5   | 28      | 224  | 40      | 242  |         |      |         |      |
| 4   | 8    | 1   |         |      | 20      | 220  | 33      | 234  | 44      | 228  |

The model in matrix notation is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{v} + \mathbf{Z}\mathbf{a} + \mathbf{Z}\mathbf{p} + \mathbf{e},$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & 22 & 484 \\ 1 & 30 & 900 \\ 1 & 28 & 784 \\ 1 & 34 & 1156 \\ 1 & 42 & 1764 \\ 1 & 40 & 1600 \\ 1 & 20 & 400 \\ 1 & 47 & 2209 \\ 1 & 55 & 3025 \\ 1 & 33 & 1089 \\ 1 & 66 & 4356 \\ 1 & 44 & 1936 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 224 \\ 244 \\ 224 \\ 236 \\ 247 \\ 242 \\ 220 \\ 239 \\ 241 \\ 234 \\ 244 \\ 228 \end{pmatrix},$$

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and for animal 1,

$$\mathbf{Z} = \begin{pmatrix} 0.7071 & -1.0288 & 0.8829 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.7071 & -0.4409 & -0.4832 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.7071 & 0.1960 & -0.7299 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

In order to reduce rounding errors, the covariates of age for the fixed regressions can be forced to have

a mean of approximately zero by subtracting 38 from all ages and 1642 from all ages squared.

The entire MME cannot be presented, but parts of the MME are given below.

$$\mathbf{W}'\mathbf{W} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

$$\mathbf{W}'\mathbf{X} = \begin{pmatrix} 3 & -34 & -2,758 \\ 4 & -16 & -1,648 \\ 3 & 21 & 1,397 \\ 2 & 34 & 3,008 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 5 & -1 \\ 5 & 1,995 & 166,883 \\ -1 & 166,883 & 14,415,319 \end{pmatrix},$$

$\mathbf{Z}'\mathbf{Z}$  is composed of the following four blocks of order 3, for the four animals with records;

$$\text{Animal1} \begin{pmatrix} 1.5 & -0.9006 & -0.2335 \\ -0.9006 & 1.2912 & -0.8383 \\ -0.2335 & -0.8383 & 1.5457 \end{pmatrix},$$

$$\text{Animal2} \begin{pmatrix} 2 & 0.7275 & 0.0259 \\ 0.7275 & 2.0233 & 1.3612 \\ 0.0259 & 1.3612 & 2.1815 \end{pmatrix},$$

$$\text{Animal3} \begin{pmatrix} 1 & -0.6235 & -0.4902 \\ -0.6235 & 0.5615 & 0.0648 \\ -0.4902 & 0.0648 & 0.5761 \end{pmatrix},$$

$$\text{Animal4} \begin{pmatrix} 1.5 & -1.1085 & 0.0134 \\ -1.1085 & 1.5121 & -1.2082 \\ 0.0134 & -1.2082 & 2.2687 \end{pmatrix}.$$

and  $\mathbf{Z}'\mathbf{X}$  for all animal is

$$\mathbf{Z}'\mathbf{X} = \begin{pmatrix} 2.12 & -7.78 & -761.55 \\ -1.27 & 19.99 & 1516.76 \\ -.33 & -18.76 & -1201.42 \\ 2.83 & 28.99 & 2458.59 \\ 1.03 & 46.44 & 4337.80 \\ 0.04 & 27.97 & 2979.60 \\ 1.41 & -5.66 & -636.39 \\ -.88 & 7.05 & 636.63 \\ -.69 & -2.14 & -22.46 \\ 2.12 & -12.02 & -1061.36 \\ -1.57 & 23.03 & 1684.81 \\ .02 & -24.57 & -1515.25 \end{pmatrix}.$$

The right hand sides of the MME are

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 2,823 \\ 2,070 \\ 68,064 \end{pmatrix},$$

$$\mathbf{W}'\mathbf{y} = \begin{pmatrix} 692 \\ 945 \\ 714 \\ 472 \end{pmatrix},$$

and

$$\mathbf{Z}'\mathbf{y} = \begin{pmatrix} 494.2629 \\ -287.6596 \\ -90.7117 \\ 690.1296 \\ 249.1165 \\ 7.3023 \\ 329.5086 \\ -200.1692 \\ -168.8920 \\ 482.2422 \\ -351.3606 \\ -7.8918 \end{pmatrix}.$$

The variance-covariance matrices of the additive and permanent environmental effects need to be known for BLUP. Normally, these are not well known and must be estimated simultaneously with the other effects of the model. Let

$$\mathbf{G} = \begin{pmatrix} 94.0000 & -3.8500 & 0.03098 \\ -3.8500 & 1.5000 & -0.0144 \\ 0.03098 & -0.0144 & 0.0014 \end{pmatrix},$$

and

$$\mathbf{P} = \begin{pmatrix} 63.0000 & -2.1263 & 0.0447 \\ -2.1263 & 0.5058 & -0.00486 \\ 0.0447 & -0.00486 & 0.0005 \end{pmatrix}.$$

The solutions to MME are

$$\hat{\mathbf{b}}' = (234.9797 \quad 1.4670 \quad -0.01399),$$

$$\hat{\mathbf{c}}' = \begin{pmatrix} -0.8630 \\ 1.2885 \\ 0.1443 \\ -0.5698 \end{pmatrix}.$$

Let the solutions for the animal additive genetic random regression coefficients be presented in tabular form as follows.

| Animal | $a_0$     | $a_1$     | $a_2$     |
|--------|-----------|-----------|-----------|
| 1      | -2.021529 | 0.175532  | -0.002696 |
| 2      | 5.751601  | -2.139115 | 0.025848  |
| 3      | -2.474456 | 2.554412  | -0.029269 |
| 4      | -5.376687 | -0.370873 | 0.002174  |
| 5      | -1.886714 | 1.464975  | -0.016963 |
| 6      | 3.333268  | -1.065525 | 0.013047  |
| 7      | 1.503398  | -1.081654 | 0.012555  |
| 8      | -2.948511 | 0.681643  | -0.008633 |

Similarly, the solutions for the animal permanent environmental random regression coefficients can be given in tabular form.

| Animal | $p_0$     | $p_1$    | $p_2$     |
|--------|-----------|----------|-----------|
| 1      | -0.296786 | 0.246946 | -0.002521 |
| 2      | 3.968256  | -730659  | 0.009430  |
| 3      | -0.834765 | 0.925329 | -0.008164 |
| 4      | -4.505439 | -441805  | 0.001257  |

**Ranking Animals** The problem is to rank the animals for selection purposes. If animals are ranked on the basis of  $a_0$ , then animal 2 would be the highest (if that was desirable). If ranked on the basis of  $a_1$ , then animal 3 would be the highest, and if ranked on the basis of  $a_2$ , then animal 2 would be the highest. To properly rank the animals, an EBV at different ages could be calculated, and then these could be combined with appropriate economic weights. EBVs were calculated for 24, 36, and 48 mo of age, and economic weights of 2, 1, and 0.5, respectively, for the three EBVs were used to compute a Total Economic Value (TEV), as

$$2 * \text{EBV}(24) + 1 * \text{EBV}(36) + .5 * \text{EBV}(48).$$

The Legendre polynomials for ages 24, 36, and 48 mo are given in the rows of the following matrix  $\mathbf{L}$ ,

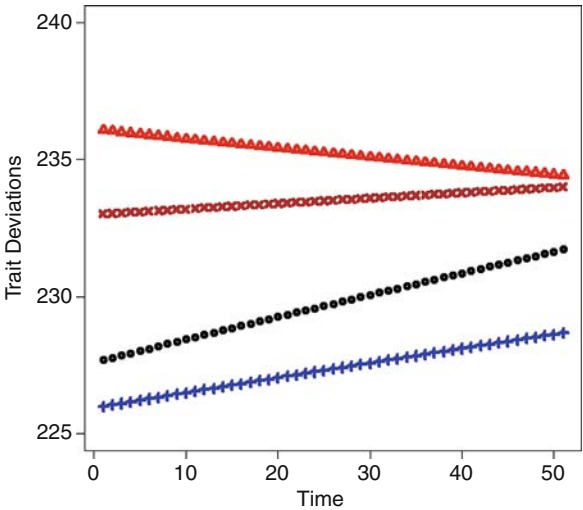
$$\mathbf{L} = \begin{pmatrix} 0.7071 & -0.8328 & 0.3061 \\ 0.7071 & -0.3429 & -0.6046 \\ 0.7071 & 0.2449 & -0.6957 \end{pmatrix}.$$

The results are shown in the following table.

| Animal | EBV(24) | EBV(36) | EBV(48) | TEV    |
|--------|---------|---------|---------|--------|
| 1      | −1.58   | −1.49   | −1.38   | −5.33  |
| 2      | 5.86    | 4.78    | 3.53    | 18.26  |
| 3      | −3.89   | −2.61   | −1.10   | −10.93 |
| 4      | −3.49   | −3.68   | −3.89   | −12.61 |
| 5      | −2.56   | −1.83   | −.96    | −7.43  |
| 6      | 3.25    | 2.71    | 2.09    | 10.25  |
| 7      | 1.97    | 1.43    | .79     | 5.76   |
| 8      | −2.66   | −2.31   | −1.91   | −8.58  |

The animal with the highest TEV was animal 2. All animals ranked rather similarly at each age on their EBVs. Rankings of animals could change with age. Thus, the pattern of growth could be changed.

**Plotting Curves** The shapes of curves of individual animals could also be plotted from the solutions, and shown as deviations from the average curves. Figure 5 contains animals 1 (black), 2 (red), 4 (blue), and 6 (brown)



Animal Breeding, Modeling in. Figure 5  
Curves for animals 1 (black), 2 (red), 4 (blue), and 6 (brown), in the example data

(brown). Animal 2 had the highest TEV, but that ranking was achieved by Animal 2 having the highest initial deviation. However, the curve declines very rapidly after time 18. The curves of the other animals do not decline as rapidly, and the curve for Animal 1 actually increases slightly. Depending on the biology of the trait and the welfare of the animal, one may decide to select for animals with increasing curves rather than declining curves. The weights in the TEV could be adjusted accordingly.

Comments

The orders of the Legendre polynomials do not need to be equal for both the animal additive genetic and animal permanent environmental effects. The order of the fixed regressions could also be different, and as seen in this example can be based on a different function of time covariates.

There are other kinds of orthogonal functions that could be used in place of Legendre polynomials, and some of these have been tried without much benefit to the analyses (Yazdi?). Another alternative has been spline (split polynomial) functions, [19, 22] in which a curve is broken down into sections. Within each section, a simple linear or quadratic function is sufficient to fit the data within that section. The sections are joined together by “knots,” the locations of which need to be estimated. Spline functions have gained some popularity lately.

Because random regression models were new to animal breeding in 1994, [20] they were applied to many different types of research problems. RRM have been most successful in the analysis of dairy cattle test day production within lactations, and in growth of pigs, sheep, rainbow trout, and beef cattle. One limitation with growth data is that animals are not usually weighed more than three or four times in their life due to the amount of labor involved in collecting weights and the stress induced on the animals during the weighing process. The orders of the regression coefficients for growth traits are usually limited to 2 or 3.

Multiple Trait Models

Animals are observed for many traits relating to production, reproduction, conformation, longevity or

fitness, and health such that knowing the total economic merit of an animal helps to keep costs of production to a minimum. Most traits are genetically correlated to each other, meaning that some genes affect more than one trait. Because contemporary animals make their records in the same environment, environmental correlations due to management, feed, and temperature also exist to affect observations on all traits. Thus, a sensible approach is to analyze groups of traits using *multiple trait models* [23–26]. In this way, information from correlated traits can be used to improve the accuracy of all trait evaluations. A multiple trait (MT) model is one in which two or more traits are analyzed simultaneously in order to take advantage of genetic and environmental correlations between traits.

**Low Heritability Traits:** MT models are useful for traits where the differences between genetic and residual correlations are large (e.g., greater than 0.5 difference) or where one trait has a much higher heritability than the other traits. EBVs for traits with low heritability tend to gain more in accuracy than EBVs for traits with high heritability, although all traits benefit to some degree from the simultaneous analysis(24).

**Culling:** Another use of MT models is for traits that occur at different times in the life of the animal such that culling of animals results in fewer observations on animals for traits that occur later in life compared to those at the start. Consequently, animals which have observations later in life tend to have been selected based on their performance for earlier traits. Thus, analysis of later life traits by themselves could suffer from the effects of culling bias, and the resulting EBV could lead to errors in selecting future parents. MT analyses have been shown to partially account, to some degree, for the selection that has taken place [25, 26]. If the percentage of missing trait observations is high and the missing observations are not due to random chance, then biases could be very large in EBVs for that trait and maybe others.

The success of an MT analysis relies on the accuracy of the genetic and residual correlations that are assumed. Computations for MT models are more complicated than for single trait analyses. If  $m$  is the number of traits and  $N$  is the number of animals in the pedigree file, then there are at least  $mN$  equations to be

solved rather than just  $N$ , and the same relative difference in memory space to hold solutions. However, computer hardware advances (with gigabytes of memory) have been very rapid in the last decade. Consequently, many genetic evaluation systems are now multiple trait systems.

## Models and Traits

Usually, traits are grouped together for MT analyses because they are observed at approximately the same point in time. For example, milk, fat, and protein yields of dairy animals are observed on the same day, per animal. Thus, for a group of similar traits, the linear models associated with those traits include the same major factors like years, ages, contemporary groups, and animal additive genetic effects. Another grouping of traits would be those for reproductive traits, observed primarily on females, such as conception rate, litter size, birthing ease, and offspring losses at birth. The linear models used for production traits are quite different from those used for reproductive traits, or health traits.

Even within a group, the linear models for the traits within a group could be different from each other. Consider two traits with a single observation per trait on animals. Let the model equation for trait 1 be

$$y_{1ij} = B_{1i} + a_{1j} + e_{1ij},$$

where  $B_{1i}$  is a fixed effect with  $p_B$  levels,  $a_{1j}$  is a random, animal additive genetic effect for trait 1, and  $e_{1ij}$  is a random residual environmental effect for trait 1.

The model equation for trait 2 might be

$$y_{2ij} = C_{2i} + a_{2j} + e_{2ij},$$

where  $C_{2i}$  is a fixed effect (different from  $B_{1i}$  for trait 1) with  $p_C$  levels,  $a_{2j}$  is a random, animal additive genetic effect for trait 2, and  $e_{2ij}$  is a random residual environmental effect for trait 2.

For example,  $y_{1ij}$  could be birth weight, so that  $B_{1i}$  could identify animals born in the same season. Trait 2 could be yearling weights and  $C_{2i}$  could identify contemporary groups of animals of the same sex, same herd, and same rearing unit within herd.

Because the two traits will be analyzed simultaneously, variances and covariances need to be specified



for the traits together. For example, the additive genetic variance-covariance (VCV) matrix (between traits) could be written as

$$\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 15 \end{pmatrix},$$

and the residual environmental VCV matrix as

$$\mathbf{E} = \begin{pmatrix} e_{11} & e_{12} \\ e_{12} & e_{22} \end{pmatrix} = \begin{pmatrix} 10 & 5 \\ 5 & 100 \end{pmatrix}.$$

The genetic and residual correlations are, respectively,

$$\rho_g = 2/(15)^{0.5} = 0.516,$$

$$\rho_e = 5/(1000)^{0.5} = 0.158$$

with heritabilities specified as

$$h_1^2 = \frac{1}{11} = 0.0909,$$

and

$$h_2^2 = \frac{15}{115} = 0.1304.$$

For all data, using  $\mathbf{A}$  as the additive numerator relationship matrix, then

$$\text{Var} \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}g_{11} & \mathbf{A}g_{12} \\ \mathbf{A}g_{12} & \mathbf{A}g_{22} \end{pmatrix}.$$

### Data Example

A two trait example with three factors is given in Table 6. Note that animals whose trait 1 observation was below 3.0 were not allowed to make a trait 2 observation. If the true variances and covariances are known, then this selection bias will be lessened through the multiple trait analysis, as long as the analysis includes the trait 1 records of animals with missing trait 2 observations.

There were two levels of factor B associated with trait 1, and 3 levels of factor C associated with trait 2. The models assumed are those given in the previous section including the covariance matrices that were given.

**Animal Breeding, Modeling in. Table 6** Example data for multiple trait models

| Animal | Sire | Dam | B-level | C-level | Trait 1 | Trait 2 |
|--------|------|-----|---------|---------|---------|---------|
| 1      | 0    | 0   | 1       | 1       | 2.3     |         |
| 2      | 0    | 0   | 1       | 2       | 2.6     |         |
| 3      | 0    | 0   | 1       | 3       | 9.8     | 53      |
| 4      | 0    | 0   | 1       | 1       | 4.7     | 4       |
| 5      | 0    | 0   | 1       | 2       | 5.5     | 63      |
| 6      | 1    | 3   | 2       | 3       | 2.5     |         |
| 7      | 1    | 4   | 2       | 2       | 8.4     | 35      |
| 8      | 1    | 5   | 2       | 3       | 8.2     | 41      |
| 9      | 2    | 3   | 2       | 1       | 9.0     | 27      |
| 10     | 2    | 4   | 2       | 1       | 7.8     | 32      |
| 11     | 2    | 5   | 2       | 2       | 2.8     |         |
| 12     | 6    | 10  | 2       | 3       | 7.4     | 67      |

### Mixed Model Equations

The models are written in matrix notation below.

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{pmatrix} \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix}$$

Let  $\mathbf{R}$  be the covariance matrix of residual effects. If observations are ordered by traits within animals, then  $\mathbf{R}$  is the direct sum of submatrices, one for each animal. The submatrices depend on which traits are observed for a given animal. If an animal has been observed for both (all) traits, then the diagonal block is  $\mathbf{E}$  as given earlier. If an animal has been observed only for trait 1, then the block is

$$\mathbf{E}_1 = \begin{pmatrix} e_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

and if the animal has been observed only for trait 2, then

$$\mathbf{E}_2 = \begin{pmatrix} 0 & 0 \\ 0 & e_{22} \end{pmatrix}.$$

With more than two traits, the number of combinations of missing traits increases. The covariance

matrix of residual effects for the example data is a block diagonal matrix,

$$\mathbf{R} = \text{diag}(\mathbf{E}_1 \mathbf{E}_1 \mathbf{E} \mathbf{E} \mathbf{E}_1 \mathbf{E} \mathbf{E} \mathbf{E} \mathbf{E}_1 \mathbf{E}).$$

The inverse of  $\mathbf{R}$  is needed for the mixed model equations, and this can be obtained by inverting the block diagonal matrices (ignoring the zero rows and columns when there are missing observations).

The resulting mixed model equations were of order 29 by 29. The solutions, for this example, were

$$B_{11} = 5.0209$$

$$B_{12} = 6.5592$$

$$C_{21} = 20.0882$$

$$C_{22} = 49.0575$$

$$C_{23} = 51.9553$$

| Animal | Sire | Dam | Trait 1 | Trait 2 |
|--------|------|-----|---------|---------|
| 1      | 0    | 0   | −0.3573 | −1.6772 |
| 2      | 0    | 0   | −0.0730 | 1.0418  |
| 3      | 0    | 0   | 0.4105  | 1.1707  |
| 4      | 0    | 0   | −0.0449 | −1.4922 |
| 5      | 0    | 0   | 0.0646  | 0.9570  |
| 6      | 1    | 3   | −0.1033 | −1,410  |
| 7      | 1    | 4   | −0.1975 | −2.2983 |
| 8      | 1    | 5   | −0.1410 | −9,633  |
| 9      | 2    | 3   | 0.3079  | 1.6227  |
| 10     | 2    | 4   | 0.1426  | 1.1273  |
| 11     | 2    | 5   | −0.1830 | 0.6418  |
| 12     | 6    | 10  | 0.1554  | 1.5089  |

Notice that every animal has an EBV for both traits, even though trait 2 was missing for some animals. The EBV for an animal with a missing trait observation is constructed based on the genetic and environmental correlations between traits, and based on their genetic relationships to other animals that were observed for both traits.

### Economic Indexes

Many producers are so overwhelmed by the numerous EBV available for so many traits that they are unable to

visualize the relative contributions of different traits to their overall productivity. Thus, economic indexes were developed where trait EBVs are scaled to unit variances, and then relative economic values are applied to the scaled EBVs. Sometimes different types of indexes are derived depending on the market objectives of the producers. A beef producer, for example, may be producing animals of high carcass merit, and would use an index that puts more weight on carcass traits of marbling score, tenderness, and dressing percentage. Another producer may be producing calves and may want more weight on the reproductive ability of cows and calf survival from birth to weaning. A particular breed may wish to be known as the “Have Everything” breed and would utilize a different set of economic weights where many traits are emphasized.

Economic weights are dynamic in that they change over time, and can often change very dramatically and quickly, usually more quickly than a breeding program can be changed due to the long generation intervals of the species involved. Producers must therefore continually reassess their goals and objectives.

Economic indexes based on multiple trait EBVs tend to be more stable over time compared to the same EBVs computed on each trait separately.

### Models for Categorical Data

Categorical data arise in animal breeding in situations where a trait is subjectively scored by the producer or a trained individual. For example, calves can be born with absolutely no assistance of the producer or to the other extreme where the producer must ask for veterinarian assistance to perform a caesarian section to deliver the calf and save the cow. Between those two extremes are different levels of difficulty of calving. Recording programs provide four or five categories of difficulty, and the producer must decide to which category a calving belongs. The assumption is that one producer assigns all calvings based on the same criteria, but a different producer may have slightly different criteria.

The number of calvings falling into each category depends on the subjectivity of the producers. In most breeds, the category for unassisted or easy births is usually very high (from 50% to 90% of all calvings), while the other categories are often much smaller and

the category for caesarian births being the smallest. The categories are ordered from one extreme to the other.

In dairy cattle, cows are scored for 30 or more conformation traits (i.e., style points) by trained judges, and each trait has 9 or more categories. Judges are trained and updated annually, but judges can differ in their abilities to score traits, particularly when they are trained to score a cow in a few minutes.

## Theory

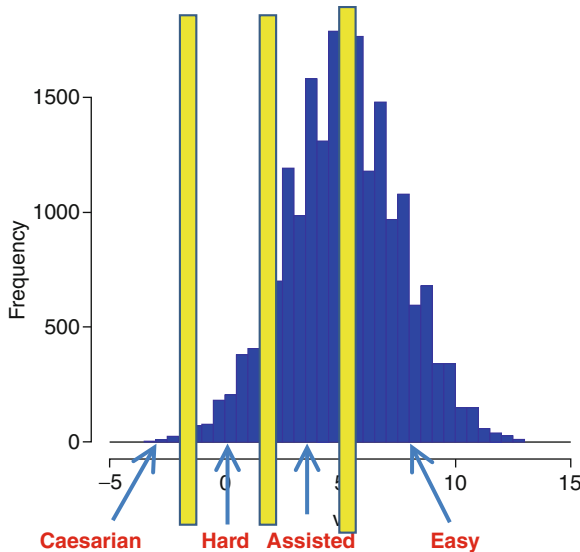
Although a trait may be scored into one of a limited number of categories, an underlying non-observable, normally distributed trait could be hypothesized [27–29]. Then thresholds along the scale are where the categories of the observed scores are defined, as shown in Fig. 6.

Thus, when an animal's underlying scale trait exceeds a threshold value, then it belongs to the next higher category. The threshold model is ideal for analyzing categorical data. The analysis follows Bayesian concepts and is nonlinear in the solutions to this model.

Write a model for the underlying scale variable.

$$\ell_{ijkm} = t_i + b_j + a_k + e_{ijkm},$$

where  $\ell_{ijkm}$  is an unobserved value on the underlying scale for the trait of interest;  $t_i$  is one of the threshold



**Animal Breeding, Modeling in. Figure 6**  
Normal distribution with thresholds for calving ease

points;  $b_j$  is a fixed factor such as age of the dam, breed, year, or season;  $a_k$  is an animal additive genetic effect; and  $e_{ijkm}$  is a residual error effect. The model could be much more complex than that described here, depending on the trait and situation.

Note that, instead of observations on the underlying scale, only the category to which an animal belongs is known. There are various quantities which need to be computed repeatedly in the analysis, and these are based on normal distribution functions.

1.  $\Phi(x)$  is known as the cumulative distribution function of the normal distribution. This function gives the area under the normal curve up to the value of  $x$ , for  $x$  going from minus infinity to plus infinity (the range for the normal distribution). For example, if  $x = 0.4568$ , then  $\Phi(x) = 0.6761$ , or if  $x = -0.4568$ , then  $\Phi(x) = 0.3239$ . Note that if there are  $m$  categories and  $k = m$ , then  $\Phi_k = 1$ .
2.  $\phi(x)$  is a function that gives the height of the normal curve at the value  $x$ , for a normal distribution with mean zero and variance 1. That is,

$$\phi(x) = (2\pi)^{-0.5} \exp -0.5x^2.$$

For example, if  $x = 1.0929$ , then  $\phi(x) = 0.21955$ .

3.  $P(k)$  is the probability that  $x$  from a  $N(0, 1)$  distribution is between two threshold points, or is in category  $k$ . That is,

$$P(k) = \Phi_k - \Phi_{k-1}.$$

If  $k = 1$ , then  $\Phi_{k-1} = 0$ .

Begin with phenotypic values for  $t_i$  based on a normal distribution. From the  $t_i$ , observations on the underlying scale can be “created.” Each one would have a different weight in the analysis due to the frequency of a category being observed. Then new values of  $b_j$  and  $a_k$  are calculated. Finally, new values of  $t_i$  are determined, and the process is repeated. Eventually, the process converges until “solutions” for all variables do not change. This is an overly simplified explanation. For more details, see [30].

## Solutions

Using the estimates from the nonlinear system of equations, the probability of an animal's offspring falling into each category can be calculated. Let  $\hat{a}_k = 0.123$ , and let

there be a 3 category trait. The solutions for the two thresholds were  $\hat{t}_1 = 0.376$  and  $\hat{t}_2 = 1.012$ . Then the value on the underlying scale for the first category would be

$$\begin{aligned} x &= \hat{t}_1 + \hat{a}_k, \\ &= 0.376 + 0.123, \\ &= 0.499. \end{aligned}$$

Then

$$\Phi(x) = \Phi(0.499) = 0.691.$$

Similarly, the probability of the animal's offspring to be in categories 1 or 2 would be based on the second threshold,

$$x = 1.012 + 0.123 = 1.135,$$

or  $\Phi(x) = 0.872$ . Thus, the proportion of offspring that would fall in category 2 would be  $0.872 - 0.691 = 0.181$ . The proportion that would be in category 3 would be  $1.0 - 0.872 = 0.128$ .

Animals could be ranked on their  $\hat{a}_i$ , or the result could be expressed as a probability of being in a particular category. For a trait like calving ease, for example, one might want to maximize the probability of having an easy birth.

## Comments

While a threshold model is a theoretically best approach to the analysis of categorical data, research has found small differences in accuracy of ranking animals from procedures that treat the category numbers as any continuously distributed trait [31]. As the number of categories increases, the differences between a threshold model analysis and simple linear model analysis become smaller.

Computational problems may arise with threshold models due to small numbers of observations in one or more categories. This often causes two categories to be merged into one to bypass the problem.

Categorical traits are often standardized to a normal scale, and then analyzed with a usual linear model [32]. If the threshold values on a normal scale were 0.376 and 1.012 for a three category trait, then all observations in category 1 would receive a score of  $(0.376 - 0)/2 + 0 = 0.188$ . Category 2 observations would be scored  $(1.012 - 0.376)/2 + 0.376 = 0.694$ ,

and category 3 would be scored  $(3 - 1.012)/2 + 1.012 = 2.006$ . Standardization is often conducted within years, or time periods in which the subjective scoring was applied consistently.

Because categorical traits are subjectively recorded, the subjectiveness can fluctuate over time. An animal that was assessed to be category 1 20 years ago, may now (if the animal were alive today at the same age) belong to category 2. The standards of the observer(s) have changed over those 20 years. The change is similar to the change in the value of the US dollar between now and 20 years ago. Today \$1 does not buy as much as \$1 back in 1990. The consequence of shifting standards (in animal breeding) is that genetic trends cannot be properly estimated. Comparing animals many years apart is not statistically accurate for categorical data. This is not a major problem because producers mainly want to compare animals that are alive today, and for this purpose, the assumption is that subjective standards change rather slowly and not by very much in a short time span of about 5 years.

Methods have also been developed for multiple trait models involving traits that are categorical and other traits that are continuous [33, 34]. These have primarily been cases of binomial traits, such as disease traits (yes or no situations). Binomial traits have only one threshold to be estimated, and binomial traits seem to benefit from a threshold model.

## Models for Survival Data

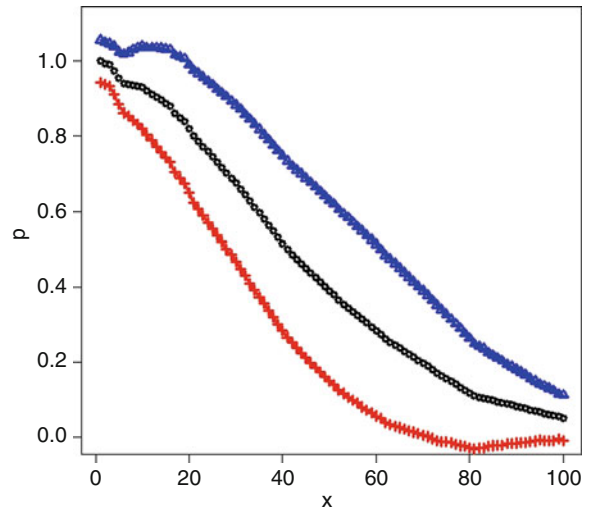
In animal production systems, animals remain in the herd or flock as long as they are productive and generate more income than expenses. Eventually, animals leave the production unit due to natural death or injury, or due to lowered production levels or reproductive problems such that the animal is deemed unprofitable. The latter factors are determined by the owner, and owners differ widely in their management skills, accounting, and decision-making abilities. There seems to be a small amount of genetic variability among animals in survival rates that do not depend on productive abilities, and genetic evaluation systems for survivability or longevity have been developed. The heritability of survival is usually less than 0.02. Dairy sires, for example, need thousands of progeny in order to have an accurate Estimated Breeding Value for survival ability of their progeny.

Survival has been defined in many different ways. In a binary sense, alive is equal to 1, every day until the animal dies and thereafter becomes 0 for the remainder of the observable time period. For example, in dairy cattle, the observable period is from first calving until 100 months later (approximately 10 years of age) when the majority of cows have died. The time intervals in that range can be hours, days, weeks, months, or years. Define a survival vector for the  $i$ th cow as  $\mathbf{s}_i$ , in terms of months (from 1 to 100), such that the values of  $\mathbf{s}_i$  are equal to 1 for every month the cow was alive, and equal to 0 for the month in which it dies and every month thereafter to the limit of 100 months after first calving.

A cow at 44 months after first calving today is still alive, and its future death time is unknown, then the values for months 45–100 need to be considered missing or not observed. For a cow that is still alive 100 months after first calving, then  $\mathbf{s}_i$  contains all ones. If a cow has gone missing from the data, in the sense that it was sold to another owner and that owner is not on a milk recording program, then the cow has not died, but the exact time of its death is not known. In this case,  $\mathbf{s}_i$  will contain 1s up until the time it was sold, and all remaining values in  $\mathbf{s}$  are unknown (neither 1 nor 0).

Having defined  $\mathbf{s}_i$  for cow  $i$ , then let  $\mathbf{S}$  be the average of all  $\mathbf{s}_i$  for cows that do not have any missing values. The elements of  $\mathbf{S}$  are equal to the probabilities of being alive at each month after first calving, and can be plotted as shown in Fig. 7. The value at a given time (month) is a probability,  $p_t$ . The black points refer to the average  $\mathbf{S}$  for all cows, while the blue and red lines indicate 2 standard deviations above and below the average, respectively. Note that the variation is less around  $p_1$  and  $p_{100}$  than in the middle of the range.

Because a curve is involved, a random regression model could be applied [35, 36] such that a separate set of curve parameters could be estimated for each animal. The model would include the year-season of first calving (with separate curves for each year-season), random herd within year-season of first calving, a variable that indicates whether the herd was increasing, decreasing, or being stable in size at a particular instance of time, a regression on the genetic EBV of an animal for production, random animal additive genetic regression coefficients, and random animal permanent environmental regression coefficients. The model would also need to account for different residual



**Animal Breeding, Modeling in. Figure 7**

Survival curve for Jersey cattle after first calving, in months

variances for each month after first calving. Meuwissen et al. [36] have shown links between various models for survival analyses. Jamrozik et al. [37] did a comparison of the effectiveness of different models for survival.

Another model is the proportional hazard model (PHM) [38],

$$\lambda(t) = \lambda_{0,s}(t) \exp[\mathbf{x}'_m \boldsymbol{\beta} + \mathbf{z}'_m \mathbf{u}],$$

where  $\lambda(t)$  is the probability of a cow being culled at time  $t$  given she was alive before time  $t$ ;  $\lambda_{0,s}(t)$  is the Weibull baseline hazard function;  $\boldsymbol{\beta}$  contains time-dependent covariates affecting the hazard function with  $\mathbf{x}'_m$  being the corresponding design vectors;  $\mathbf{u}$  is a vector of random factors including herd, year-seasons, and animal additive genetic effects; and  $\mathbf{z}'_m$  is the associated incidence vector. Thus, a nonlinear system of equations needs to be solved.

Lastly, survival could be defined as a discrete trait, such as the survival of an animal to the end of first, second, or third lactation. The three “traits” can be analyzed as a multiple trait system, in which cows need not be observed for all traits. This is similar to the random regression approach except the number of months after first calving is reduced to just 3 or 4 broader periods of time instead of monthly.

Because the heritability of survival is so low, and analyses are affected by the animals that are still alive at the time of analysis, perhaps a better indicator of



survivability would be a measure of profitability of the animal. The traits that are part of profitability would tend to have higher heritabilities, and would contribute toward an animal remaining in the herd. Further research on this area is needed.

## Added Complexities

### Heterogeneous Variances

In dairy cattle, the within contemporary group variation of records was deemed to differ between contemporary groups. That implied that the contemporary group effects were being sampled from different populations. Consequently, both the genetic and residual variation could be allowed to differ between contemporary groups. Bayesian methods were developed to simultaneously compute solutions to mixed model equations and to estimate the necessary contemporary group variances. Often, only the phenotypic variances within contemporary groups were assumed to differ such that the heritability was constant across contemporary groups.

### Robust Estimation

Related to heterogeneous variances was the problem of outliers, records that were extreme values in the distribution of phenotypes. Sometimes the extremes were deliberately created by producers who thought they knew how to manipulate the data so that their good animals would receive high EBVs. Other times, the extremes were due to errors in recording, and occasionally were naturally extreme. Regardless of the reason for their existence, the result was a bias to EBVs of one animal or an entire group of animals. Robust estimation methods were introduced to pull the extreme records back toward the mean. Either the record itself or the estimated residual effect of the record could be modified toward the mean if it was beyond two and half standard deviations from the mean. Robust methods usually reduce the biases caused by extreme outliers, but the method requires nonlinear estimation.

### Count Data

In swine and sheep breeding, interest is in litter size, which is not normally distributed and also not categorical, but is known as *count* data. Another trait is number of services to attain conception, or number of ovulations.

Behavioral traits can also be count data, such as the number of times an animal takes water during the day. The appropriate distribution for count data is a Poisson distribution, and Bayesian methods for handling and modeling this distribution have been developed.

## Hierarchical Models

Another Bayesian development was that of hierarchical models where a trait is modeled in the usual way, but then a component of that model has its own separate model. Thus, two models are solved in steps, until the solutions satisfy both models.

## Model Comparisons

Methods have also been developed to compare different models, but even further to select from among a group of models for the better fitting model. Model comparisons are usually conducted using an “estimation” data set (with which to estimate variances and solutions to mixed model equations), and a second, sometimes smaller, “validation” data set to judge how well observations and rankings of animals can be predicted.

## International Comparisons

Some species of livestock are traded among countries, depending on health restrictions, and this necessitates comparing the genetic abilities of animals between countries. Usually the genetic evaluation of animals within a country has enough complexities to it such that the models used in country A are very different and incompatible with the models in country B. Thus, EBVs from the different countries are collated and a multiple trait model (where each country is a different trait) is applied to the EBVs of male animals. Genetic correlations among countries are less than unity and rankings of animals may differ between countries. The Interbull organization in Uppsala, Sweden routinely computes evaluations for dairy bulls from 23 or more countries every year.

## Future Directions

Molecular genetics will dominate animal breeding research in the next 20 years. The discovery of millions of single nucleotide polymorphisms as genetic markers has forced animal breeders to restudy their basic

quantitative genetics notes. The DNA genomes of animal species are being completely mapped, and someday soon all of the genes and their locations in the genome will be known. This is an exciting time for animal breeders and the opportunities for research will be huge.

Already DNA markers are being used to determine the relatedness of individuals to each other, to measure the amount of genetic diversity in a species, to identify genes with major effects on production traits, and to select future breeding animals more accurately than previously possible and as soon as an animal is born. Models for genetic evaluation of animals have been modified to include genetic marker information and are thus becoming more complicated, involving many thousands of markers such that computing issues are more demanding than for multiple trait models. The result is that animals are being selected more intensely at an early age, and the samples of progeny are no longer a random group of all possible progeny of a particular mating, but are highly selected. This lack of randomness of sampling progeny may cause biases in current genetic evaluation models. The possibility also exists that rates of inbreeding could be increased. Modeling of production systems and strategies for making continued genetic change will be utilized.

Bayesian methods will likely be utilized more heavily than the methods of Henderson [2] in order to deal with DNA modifications to statistical models. Structural equation models [39] will try to determine the cause and effects of multiple traits. Statistical models, linear and nonlinear, continue to be required background training for animal breeders. The complexity and usefulness of models will grow even with the emergence of genomics.

## Bibliography

1. Lush JL (1945) Animal breeding plans. Iowa State University Press, Ames
2. Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Ontario
3. Gianola D, Fernando RL (1986) Bayesian methods in animal breeding. J Anim Sci 63:217–244
4. Sorensen D, Gianola D (2002) *Likelihood*, Bayesian and MCMC methods in quantitative genetics. Springer, New York
5. Johannsen WJ (1909) Elemente der Exakten Erblchkeitslehre. Gustav Fischer, Jena
6. Wright SE (1933) Evolution and the genetics of populations, vol I. University of Chicago Press, Chicago
7. Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used for prediction of breeding values. Biometrics 32:69–79
8. Meuwissen THE, Luo Z (1992) Computing inbreeding coefficients in large populations. Genet Sel Evol 24:305–313
9. Quaas RL (1988) Additive genetic model with groups and relationships. J Dairy Sci 71:1338–1345
10. Robinson GK (1986) Group effects and computing strategies for models for estimating breeding values. J Dairy Sci 69:3106–3111
11. Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. Biometrika 58:545–554
12. Graser H-U, Smith SP, Tier B (1987) A derivative free approach for estimating variance components in animal models by restricted maximum likelihood. J Anim Sci 64:1362–1370
13. Meyer K, Smith SP (1996) Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. Genet Sel Evol 28:23–49
14. Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet Sel Evol 25:41–62
15. Bijma P (2006) Estimating maternal genetic effects in livestock. J Anim Sci 84:800–806
16. Thompson R (1976) The estimation of maternal genetic variances. Biometrics 32:903–918
17. Schaeffer LR, Kennedy BW (1989) Effects of embryo transfer in beef cattle on genetic evaluation methodology. J Anim Sci 67:2536–2543
18. Heydarpour M, Schaeffer LR, Yazdi MH (2008) Influence of population structure on estimates of direct and maternal parameters. J Anim Breed Genet 125:89–99
19. Meyer K (2005) Random regression analyses using B-splines to model growth of Australian Angus cattle. Genet Sel Evol 37:473–500
20. Schaeffer LR, Dekkers JCM (1994) Random regressions in animal models for test-day production in dairy cattle. In: Proceedings of the 5th world congress of genetics applied to livestock production, Guelph, 7–12 Aug 1994, vol 18 p 443
21. Schaeffer LR (2004) Application of random regression models in animal breeding. Livest Prod Sci 86:35–45
22. Misztal I (2006) Properties of random regression models using linear splines. J Anim Breed Genet 123:74–80
23. Henderson CR, Quaas RL (1976) Multiple trait evaluation using relatives' records. J Anim Sci 43:1188–1197
24. Pollak EJ, Quaas RL (1981) Monte Carlo study of within-herd multiple trait evaluation of beef cattle growth traits. J Anim Sci 52:248–256
25. Schaeffer LR (1984) Sire and cow evaluation under multiple trait models. J Dairy Sci 67:1567–1580
26. Pollak EJ, van der Werf J, Quaas RL (1984) Selection bias and multiple trait evaluation. J Dairy Sci 67:1590–1596
27. Gianola D (1982) Theory and analysis of threshold characters. J Anim Sci 54:1079–1096
28. Harville DA, Mee RW (1984) A mixed model procedure for analyzing ordered categorical data. Biometrics 40:393–408

29. Foulley JL, Gianola D (1984) Estimation of genetic merit from bivariate all or none responses. *Genet Sel Evol* 16:285–306
30. Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical data with a threshold model. *Genet Sel Evol* 15:201–223
31. Meijering A, Gianola D (1985) Linear versus nonlinear methods of sire evaluation for categorical traits: a simulation study. *Genet Sel Evol* 17:115–132
32. Snell EJ (1964) A scaling procedure for ordered categorical data. *Biometrics* 20:592–607
33. Foulley JL, Gianola D (1986) Sire evaluation for multiple binary responses when information is missing on some traits. *J Dairy Sci* 69:2681–2695
34. Simianer H, Schaeffer LR (1989) Estimation of covariance components between one continuous and one binary trait. *Genet Sel Evol* 21:303–315
35. Veerkamp RF, Brotherstone S, Engel B, Meuwissen THE (2001) Analysis of censored survival data using random regression models. *Anim Sci* 72:1–10
36. Meuwissen THE, Veerkamp RF, Engel B, Brotherstone S (2002) Single and multitrait estimates of breeding values for survival using sire and animal models. *Anim Sci* 75:15–24
37. Jamrozik J, Fatehi J, Schaeffer LR (2008) Comparison of models for genetic evaluation of survival traits in dairy cattle: a simulation study. *J Anim Breed Genet* 125:75–83
38. Ducrocq VP, Solkner J (1994) “The survival kit”, a FORTRAN package for the analysis of survival data. In: *Proceedings of the fifth world congress on genetics applied to livestock production*, Guelph, 7–12 Aug 1994, vol 22, pp 51–52
39. Wu XL, Heringstad B, Gianola D (2010) Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. *J Anim Breed Genet* 127:3–15

## Animal Genetic in Environment Interaction

ERLING STRANDBERG

Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

### Article Outline

Glossary

Definition of the Subject

Introduction

Plasticity, Environmental Sensitivity, Reaction Norms and Genotype by Environment Interaction

Models to Describe Genotype by Environment Interaction

Consequences for Breeding Programs

Future Directions

Bibliography

### Glossary

**Genotype by environment interaction** It exists when the difference between the phenotypic values of two genotypes is not the same in two environments.

**Plasticity** The ability of an individual to respond to changes in the environment.

**Reaction norms** The phenotypic expression of a genotype as a function of the environment.

### Definition of the Subject

The existence of genotype by environment interaction ( $G \times E$ ) makes animal breeding more complicated. It means that the same genotype is not the best in all environments, and it implies that separate breeding programs might be needed to cater for these different environments. However, separate (and therefore, smaller) breeding programs might be less efficient than one large program. Small breeding programs might also encounter problems with inbreeding depression, but on the other hand, several populations with different breeding programs and breeding goals might increase the overall genetic diversity. Therefore,  $G \times E$  is an important factor to consider when creating breeding programs for animals, especially in a global setting.

### Introduction

The ability to respond to changes in the environment is a vital characteristic of all organisms. This ability is called *phenotypic plasticity* or sometimes, *environmental sensitivity*. Genetic variation in plasticity will lead to *genotype by environment interaction*. This paper starts by describing the phenomenon of plasticity, a term which is not well known in animal breeding. To describe it, several situations will be illustrated, with and without plasticity, by use of reaction norms. This will be followed by a description of various statistical models that can be used to study  $G \times E$ , including a brief description of genetic heterogeneity of residual variance. Finally, some consequences of  $G \times E$  for breeding programs will be discussed.

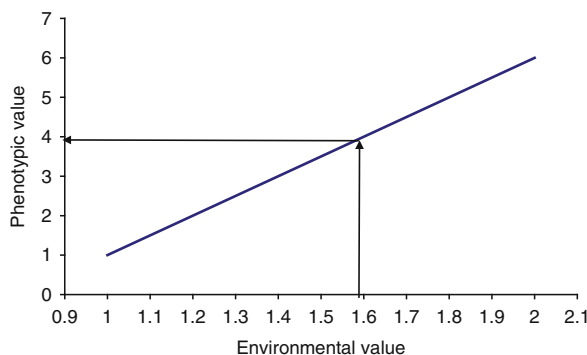
### Plasticity, Environmental Sensitivity, Reaction Norms, and Genotype by Environment Interaction

A *reaction norm* describes the phenotypic expression of a genotype as a function of the environment. One can say that the reaction norm translates the environmental values into phenotypic values. In Fig. 1, the reaction norm for a genotype is shown. The horizontal  $x$ -axis describes the environment, for simplicity let's assume a continuous scale, e.g., ambient temperature for a certain organism. The vertical  $y$ -axis gives the phenotypic value for this genotype for each environmental value. In this example, the reaction norm is linear, but it could have any form.

In Fig. 2a, genotypes that show no plasticity are depicted. Note that there is still variation in the level among genotypes, so there is genetic variation in the trait. In Fig. 2b, genotypes showing plasticity are presented. However, all genotypes react in the same way to an environmental change, i.e., there is no variation in plasticity.

In Fig. 2c, the genotypes are plastic, but some are more plastic than others. In other words, there is also variation in plasticity. Now, the genotype with the highest phenotypic value in the low environment also has the highest value in the highest environment. If high phenotypic value is desirable, which genotype to choose would be indisputable.

In Fig. 2d, there is also variation in plasticity; however, here the reaction norms also cross. Which genotype is the best, now depends on in which environment this genotype should be used.



**Animal Genetic in Environment Interaction. Figure 1**  
Schematic description of a reaction norm for a genotype

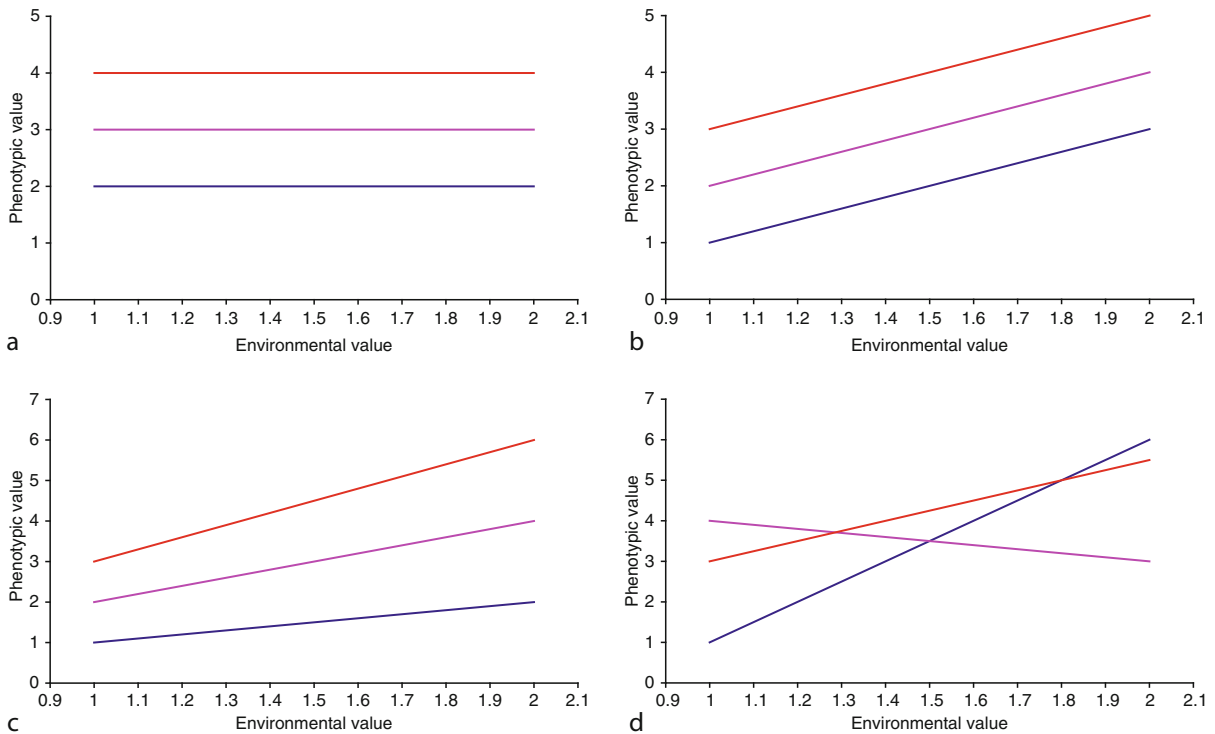
Thus, differences in phenotypic plasticity between genotypes result in *genotype by environment interaction*. Strictly speaking,  $G \times E$  means that the difference between the phenotypic values of two genotypes is not the same in two environments. If the difference changes sign between environments, there is re-ranking of genotypes (Fig. 2d). If the difference changes in size only, there is a scaling effect (Fig. 2c). Note, however, that if the environmental scale had been drawn further to the left in Fig. 2c, there would have been re-ranking also here.

### Comparison with the Usual Genetic Model

The most common quantitative genetic model is  $P = G + E$  [1]. In this description,  $G$  is defined as the *genotypic value* and  $E$  as the *environmental deviation*. “We may think of the genotype conferring a certain value on the individual and the environment causing a deviation from this, in one direction or the other” [1]. This might already from the beginning give the impression that  $G$  is the central factor, whereas  $E$  is just a nuisance, the contribution of which should be decreased as much as possible.

This standpoint actually makes sense from a traditional animal production point of view. The intention is to give all animals as good and standardized treatment as possible in order for them to produce in the best possible way. Therefore, as little as possible should be out of human control ( $E$ ). Then it can be considered that all animals have the same macro-environment (same production system, same kind of feed, etc.) and are only affected by microenvironmental effects. To the microenvironmental effects, one might count the effect of the individual farmer (within the rather standardized system), season of production or parturition, etc. These factors can be adjusted for, if known. Thus the environmental or residual variation ( $\sigma_E^2$ ) can be decreased. However, there are always some unknown microenvironmental effects (e.g., that the cow is standing close to a door with a cold draft during winter). This will end up in the residual. Therefore, given a certain phenotypic variance, the more that can be adjusted away, the higher the heritability of the trait, and the higher the expected selection response.

Expressed in a graphic way, this traditional model ( $P = G + E$ ) is described in Fig. 2a, the genotypic value



**Animal Genetic in Environment Interaction. Figure 2**

Description of the reaction norms for three genotypes. **(a)** No plasticity but genetic variation in level. **(b)** Plasticity but no variation in plasticity. **(c)** Variation in plasticity but no re-ranking of genotypes. **(d)** Variation in plasticity and re-ranking of genotypes

only gives a shift in the level of the curve. There may be environmental effects affecting the phenotypic value (as in Fig. 2b), but they affect all individuals in the same way. After adjusting for them (e.g., by a regression), Fig. 2a still applies.

If there are sufficiently different environments (e.g., Northern hemisphere vs. tropical environments), one would still be able to use the traditional model, or rather, two of the traditional models. One would assume that the trait expressed in the tropics is actually a different trait genetically from that in, say, Europe. However, within each environment  $i$ , the model  $P_i = G_i + E_i$  would still apply.

The difference between this viewpoint and that embodied in the reaction norm and phenotypic plasticity approach is quite substantial, certainly in a conceptual way. In the reaction norm approach, *the environment and the genotype are on equal standing* – there is no way a phenotype can appear without an

environment. Philosophically, this makes a lot of sense; a genotype without its environment is nothing, or at least not a phenotype. Certainly, the environment is not something to be “adjusted away” [2, 3].

Staying on the rather philosophical level, it might even be somewhat hard to define what “the environment” is. Obviously, the external environment will qualify (temperature, climate, amount and quality of food etc.). However, the internal environment within an organism is also an environment for the genome. If Richard Dawkins were asked, the whole organism (except the genome) could be considered as the environment (for the genes) [4]. In animal breeding, these fine distinctions are usually disregarded; it is difficult enough to define the external environment!

As stated earlier, the  $P = G + E$  model can be said to make sense from an animal (and plant) breeding perspective, at least given certain conditions. It is clear why this model was developed within this setting.



The same applies to the reaction norm model and the developments related to phenotypic plasticity. These models were developed mainly within evolutionary biology and genetics, where the focus is on organisms living under natural conditions. Here, there is no way to standardize the environment – the environment is what it is and it is up to the organism to adapt (or die). Populations may adapt (to a changed or variable environment) by changing genetically, i.e., individuals with higher fitness give rise to more offspring that in turn survive better and so on. After some generations, the population has a different genetic constitution (gene frequencies have changed), which is better suited to this environment, and fitness is higher than before [2].

However, changing genetically is not an option for a given individual. So, what an organism really needs is plasticity! If it can change its phenotype – at least somewhat – it will have a better chance of surviving. Now, it is not necessarily so, that even though plasticity would be good for the individual, that it would occur: there has been a lot of discussing in evolutionary genetics literature about when plasticity is expected to develop in natural populations, but that will not be covered here [2, 5]. Students of natural populations and evolutionary genetics from very early on discovered that *there is phenotypic plasticity* in all organisms studied, at least for some traits. Therefore, it is a reality, and it should be modeled somehow.

### Canalization, Homeostasis, and Plasticity

There are some terms that have been used in the evolutionary literature that need some explanation because they are sometimes referred to as the opposite of plasticity [2, 5].

Canalization means that a genotype can *buffer against small changes in the environment and in its own genotype* (mutations). The buffering means that even if small changes occur, the resulting phenotype is more or less the same. These terms are most commonly used for developmental traits, e.g., even if the environment is somewhat variable developmental plan is still followed and the resulting body, say, is still the same. It is, for instance, canalization that makes sure that cows have four legs whether they are born in Sweden or in Africa. In developmental genetics, if the environment

has an effect on the phenotype, it is usually an either-or effect, i.e., one type or another, and not a gradual effect.

Homeostasis is the outcome of canalization. Homeostasis measures the degree of (in)variance in the phenotype when the individual is perturbed by changes either in the environment or in the genome (by mutation). A more canalized genotype has higher homeostasis (is changed less) and shows lower variance. In his definition of canalization, Waddington [6] defined it for *minor* variations in conditions, what might be called microenvironmental variation. As an example of canalization, he used the environmentally triggered metamorphosis of axolotls (salamanders)! They produce one of two distinct phenotypes; which one is defined by a large change in environment. Once the developmental pathway is chosen, however, small variations in environment do not affect the outcome, and within each phenotypic outcome, there is canalization.

### Models to Describe Genotype by Environment Interaction

There are basically three different models to describe the extent of  $G \times E$ . For all methods, observations on the same or related individuals in two or more different environments are needed to study  $G \times E$ . In some organism (e.g., some plants), it is possible to use clones (i.e., numerous copies of the same genotype) and put them in different environments. With animals, that is generally not possible. However, the common use of artificial insemination in, e.g., dairy cattle makes it possible to compare the performance of daughters of the same sires in different environments [7].

In the following, the three methods will be described, not with the intent of understanding how to estimate  $G \times E$  using these models (which is relatively straightforward), but more to understand the interpretation of these models, and the type of  $G \times E$  that is detected.

### Interaction Term Model

The traditional genetic model (e.g., as in [1]) is usually written as:  $P = G + E$ , where the phenotype  $P$  is made up of a genotypic value and an environmental deviation (a residual term) (The mean is either assumed to be included in  $G$  or that  $P$  is expressed as a deviation from the overall mean). This model corresponds to

Fig. 2a, where the effect of the genotype is just to shift the level by a certain amount, regardless of the environment. In this terminology, the genotype by interaction is often simply written as  $P = G + E + G \times E$ . This terminology is incorrect and confusing: it doesn't make sense to have an interaction with the residual.

The model can be rewritten slightly such that the phenotypic value of an individual is described as the sum of the genotypic value, an environmental value, and a residual:

$$P = G + E + e \quad (1)$$

The environmental value  $E$  could, e.g., be classification into herds, herd production classes, production systems or countries. When interaction between genotype and environment exists an interaction component,  $G \times E$ , is added to the equation:

$$P = G + E + G \times E + e \quad (2)$$

The phenotypic variance ( $\sigma_P^2$ ) of the observed phenotypes ( $P$ ) can be derived from Eq. 2 as:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + \sigma_{GE}^2 + \sigma_e^2 \quad (3)$$

assuming all covariances being zero, and that all main effects are random ( $E$  might normally be considered as fixed, having only few levels).

If this model is compared to the graphs in Fig. 2, it is seen that in neither Fig. 2a nor 2b would this model detect any  $G \times E$  (correctly so). The phenotype in Fig. 2b can be exactly described by the sum of the genotypic values and an environmental value, which is the same for all genotypes. However, in both Fig. 2c and 2d, a nonzero  $G \times E$  term would show up, simply as a deviation from main effects model. Whether the  $G \times E$  would give rise to re-ranking or not is not directly seen from the size of the  $G \times E$  term ( $\sigma_{GE}^2$ ), however. One would have to add  $G + E + G \times E$  for each genotype and check whether there is much re-ranking for the environments chosen.

From this model, one could define the amount of  $G \times E$  as a “heritability” of plasticity as  $\sigma_{GE}^2 / \sigma_P^2$ . This measure is similar to a heritability in that it is a variance ratio; however, it is not useful to predict response to selection for plasticity in the same way as the “additive,” narrow sense heritability [2, 8–10].

There is no specific limitation to the number of environments that can be defined in the term  $E$  in

Eq. 2, e.g., this type of model has been used to estimate sire by herd interactions.

### Multiple-Trait Model

The second method used to describe  $G \times E$  is based on phenotypic values in different environments and genetic correlations ( $r_g$ ) between these. The phenotypic expression in the two environments is seen as two separate traits and  $r_g$  can be studied to see whether  $G \times E$  exists. When  $r_g$  between the phenotypic values of the same genotype expressed in different environments is high, the phenotypic expression is considered as the same trait in the different environments. In other words, if  $r_g$  between the phenotypic expressions of the trait in two different environments is close to 1, there is no  $G \times E$ . When  $r_g$  is low, the phenotypic expressions in the different environments are not the same trait and this is an indication of  $G \times E$  [1, 11].

The genetic correlation ( $r_g$ ) can be estimated using a multiple-trait analysis based on grouping herds with similar production environments to clusters and treating the observations from the different clusters as separate traits.

By just considering the genetic correlation between two environments, the  $G \times E$  of interest is that type which gives rise to re-ranking (Fig. 2d). However, one could also use the estimates of genetic variances from the multiple-trait analysis to describe the kind of  $G \times E$  that only gives rise to a scaling effect (Fig. 2c). The genetic correlation is not affected by scaling if the scaling is purely multiplicative.

This method can be used even if the environments cannot be ordered according to some meaningful scale. However, if there exists a continuous underlying scale and the environments chosen are just some representations of that scale (e.g., herd production levels: low, medium, and high), this approach can be modified to describe, by a covariance function, an infinite number of separate traits over a continuous gradient. The covariance function model was developed to model, e.g., growth trajectories, morphology, and reaction norms [12, 13]. Briefly, this method is based on applying a function to the estimated (co)variance components from a limited number of traits. Using this function, one can predict the variance for any environment (normally within the range studied in the data)

and also covariances between two environments. Further developments make it possible to estimate the covariance function directly without first using a MT approach (review in Gilmour and Thompson [14]).

### Reaction Norm Model

When the production environment can be described as a continuous variable, a third method, called the reaction norm model, is possible to use [15]. The definition of the reaction norm (RN) has already been given: the phenotypic expression of a genotype as a function of the environment [16].

In population and evolutionary genetics, this model has often been statistically analyzed using a fixed regression approach as genotypes have been placed in the different environments. However, in animal breeding, predicting breeding values as random effects is a common practice, and therefore it is natural to estimate the parameters of the RNs for each genotype from a random regression approach [17]. A simple model based on ordinary polynomials with a fixed set of regression coefficients (an average RN) and a random set for each individual is:

$$y = \mu + \sum_{f=1}^{nf-1} \beta_f x^f + \sum_{i=0}^{ni-1} b_{im} x^i + e \quad (4)$$

where  $x$  is the environmental value for the phenotypic value  $y$ ,  $\mu$  is the overall mean (fixed intercept),  $\beta_f$  are fixed regression coefficients, and  $b_{im}$  are random regression coefficients for animal (or sire),  $m$ :  $b_{0m}$  is the BV for level,  $b_{1m}$  is the BV for slope, etc.  $G \times E$  is defined by variation in linear and higher terms.

If the (co)variance matrix for the random coefficients is defined as  $\mathbf{G}$  (order  $n_i \times n_i$ ), then the genetic covariance between environments  $x_j$  and  $x_k$  can be written as  $\mathbf{x}'_j \mathbf{G} \mathbf{x}_k$ , where  $\mathbf{x}_j$  is a column vector with elements  $\{x_j^i\}$  for  $i = 0 \dots n_i - 1$ . For the linear RN, the vector  $\mathbf{x}_j = [1 \ x_j]'$ . The genetic correlation between environments  $j$  and  $k$  is  $\mathbf{x}'_j \mathbf{G} \mathbf{x}_k / \sqrt{\mathbf{x}'_j \mathbf{G} \mathbf{x}_j \mathbf{x}'_k \mathbf{G} \mathbf{x}_k}$  and the genetic variance in each environment is  $\mathbf{x}'_j \mathbf{G} \mathbf{x}_j$ . These can then be used to describe the re-ranking and scaling  $G \times E$ . The heritability in each environment can be estimated as  $h^2 = \mathbf{x}'_j \mathbf{G} \mathbf{x}_j / (\mathbf{x}'_j \mathbf{G} \mathbf{x}_j + \sigma_e^2)$ , where  $\sigma_e^2$  is the residual variance from Eq. 4.

*Using phenotypic average as environmental scale.* In the ideal situation, the environmental scale is

determined by the researcher, e.g., by subjecting the animal to various temperature, light, or nutrient conditions, and the trait measured is some phenotypic character. Although this situation may occur also in animal breeding situations, more often the environmental scale is directly related to the phenotypic values studied. The simplest example is probably the use of herd (or herd-year) average milk yield as the environmental scale when analyzing the phenotype milk yield [18, 19].

The herd average is most likely a combination of many factors, but it may give a good overall description of the type of environment the cow is exposed to, and the herd average is a practically useful description of the environment. However, the same trait occurs as both dependent and independent variable. To avoid a direct relation, the individual's own phenotypic value could be excluded from the herd average used for that individual. But a further problem is that the phenotypic herd average contains also the genetic component of those animals' phenotypes and is not only a measure of the environment. If the genetic material is not used randomly over herds, one would expect some bias to be introduced. One suggestion is to use an iterative procedure, either in a repeated REML [20] or a Gibbs sampling approach [21], in a model where the fixed regression in Eq. 4 is replaced with a herd effect. Simply expressed, this effect is estimated, used as the  $x$ -value, estimated again, and so on. It has been shown that this method gives more unbiased estimates of  $\mathbf{G}$  than using the phenotypic herd averages. The latter approach results in more uncertainty in the  $x$ -values, which gives an underestimation of variance of the slopes of the reaction norms [21].

*Group size.* The group size used as basis for the calculation of the environmental values is also of importance. Naturally, a large group size is desirable to get as precise estimate of the environment as possible. Imprecise values for the  $x$ -values in a regression analysis is expected to lead to lower regression coefficients, which in this situation means lower variance of slopes and  $G \times E$  may be underestimated [20]. However, increasing the group size (e.g., using herd average over several years rather than herd-year average) may also mean that the same environment is not measured anymore, leading to the same problem as above. These two factors must be balanced in a pragmatic way.

*Improved environmental scales.* There has been some work on defining environmental scales that are not directly dependent on the trait analyzed. Climate and weather conditions (temperature, humidity, rainfall), herd-year SD of production, replacement rate, persistency of lactation curves, calving patterns, and other measures of management practices have been studied [22–26]. This work should continue, and factor or principal component analysis might be of help in defining few but distinct environmental scales [e.g., 27]. Improved farm and animal data (e.g., management practices, feeding system, and intensity) would be of great value in this respect.

*Heterogeneous residual variance.* In model (4), it was assumed that the residual variance was constant over all environments. For many traits and environmental scales, this is not reasonable. For a sire model, it makes even less sense. By including the random regression, it is assumed that  $\frac{1}{4}$  of the genetic variance is dependent on the environment, but the remaining  $\frac{3}{4}$  (included in the residual) are not. A rough description of the residual variance structure can be found by saving the residuals from fitting model (4), dividing them into groups along the environmental scale, and calculate the residual variance within each group. If this indicates heterogeneous variance, it should be accounted for [e.g., 28].

*Multiple environmental scales.* A given phenotypic trait may be influenced by several environmental scales. One can accommodate this by extending model (4) to:

$$y = \mu + \sum_{i=1}^{ni-1} \beta_i x^i + \sum_{i=0}^{ni-1} b_{im} x^i + \sum_{j=1}^{nj-1} c_{jm} z^j + e \quad (5)$$

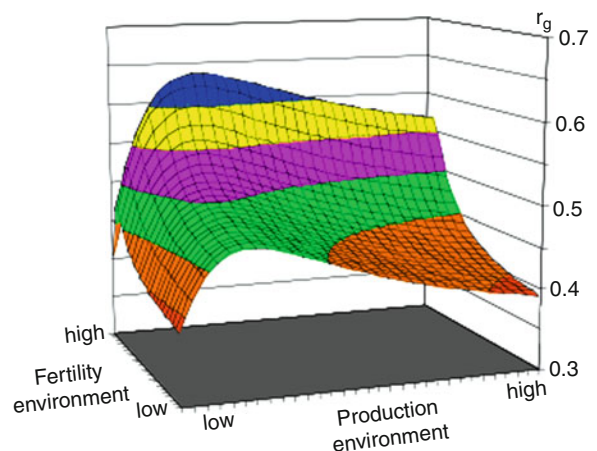
In a linear reaction norm, there is one intercept, and one slope for each environmental scale,  $x$  and  $z$ . The genetic covariance and correlation between environments can now be depicted as a function of both environmental scales, i.e., as three-dimensional graph [18, 22].

*Multiple-trait reaction norms.* Reaction norm models, as other models, can be analyzed for more than one trait at a time. The  $G$ -matrix would then contain off-diagonal sub-matrices pertaining to the covariance between the traits, and the diagonal sub-matrices would have the meaning as before.

As an example, in Fig. 3, the genetic correlation between the traits protein yield and fertility (days open) is shown [18]. It can be seen that the genetic correlation changes with the environment. Note that traits could share the same environmental scale but also have specific scales. That a certain trait shows  $G \times E$  with respect to a certain environmental scale does not mean that all traits will do so.

*Type of reaction norm function.* In the example given, ordinary polynomials were used with the origin set to the average environment. It is also recommendable to standardize the environmental scale(s); this makes (co)variance parameters easier to interpret. Another common approach is to use Legendre polynomials. These often make convergence easier than ordinary polynomials. The resulting regression coefficients are only defined within a predetermined interval (usually from minimum to maximum environmental value, redefined from  $-1$  to  $+1$ ). This means that the level is defined in the middle of this interval, which is not necessarily the average environment. For other purposes, splines have been successfully used. Regardless of the approach used, it is vital to describe exactly what was done, otherwise parameter estimates may be difficult to interpret.

*The shape of the reaction norm and the shape of the variance function are strongly connected.* For a linear

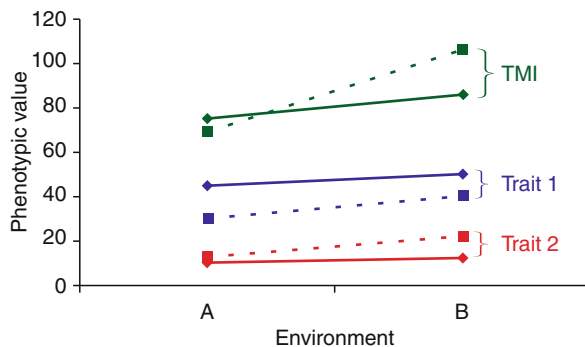


**Animal Genetic in Environment Interaction. Figure 3** Genetic correlation between protein production and fertility (days open) in dairy cattle, in relation to environmental conditions (Data from Kolmodin et al. [22])

reaction norm, the variance function becomes  $\sigma_a^2 + 2\sigma_{a,b1}x + \sigma_{b1}^2x^2$ . Because the term in front of the quadratic term is always positive, the variance curve will always be concave with an intermediate low point. In a certain data range, it may be increasing or decreasing as the minimum may be outside the range. However, it will never have a maximum intermediate point. This is a rather limiting feature of the linear reaction norm approach. With higher polynomial terms, the variance function becomes less constrained. The covariance function or character process approaches in a way work in the other direction and estimate the shape of the covariance function first. The shape of that function could be an indicator of what shape the reaction norms should be allowed to have. In most applications, only linear RNs have been found; however, there are exceptions where also higher-order RNs have been found [29–31].

*Scaling can give re-ranking.* It can be shown that even if there is only the scaling type of  $G \times E$  for the traits, there can be re-ranking in the total merit index. A simple example is given in Fig. 4. Therefore, scaling  $G \times E$  should not be considered irrelevant until the total evaluation has been done [32]; however, generally re-ranking  $G \times E$  for traits is expected to be much more important.

The indication of  $G \times E$  is that there is variation in the coefficients of the reaction norms. For linear reaction norms, this means that there is variation in the



**Animal Genetic in Environment Interaction. Figure 4** Breeding values for two individuals for two traits and total merit index (TMI) in two environments. There is scaling  $G \times E$  for both traits, but re-ranking  $G \times E$  for the total merit index (TMI =  $1 \times$  trait 1 +  $3 \times$  trait 2)

slopes. This definition would pick up plasticity of the kind shown in Fig. 2c–d. As for the interaction term model, this variation does not directly tell us how much re-ranking there is: in order to find out, one could calculate the predicted performance in various environments and get the correlation (or rank correlation) between these values.

### Genetic Heterogeneity of Residual Variation

The discussion so far has been related to macro-environmental sensitivity, i.e. the reaction of genotypes to large and known changes in the environment. However, as already mentioned, there is always unknown microenvironmental variation, variation that cannot be adjusted away or analyzed with any of the above methods, because it is not associated with any known factor. If there is genetic variation in how animals react to this variation, then there is genetic heterogeneity of residual variation (GHRV).

There are several different models possible to analyze GHRV, well summarized in Mulder [33]. The simplest additive model draws the residual not from a distribution with a constant variance but where it also depends on an additive breeding value,  $A_v$ . Thus, each individual has two breeding values, one (usual) breeding value for the mean ( $A_m$ ) and one for the residual variance ( $A_v$ ) [34]. One perhaps slight drawback with this model is that  $A_v$  is drawn from a normal distribution, and the sum of the average residual variance and  $A_v$  could become negative. The exponential model overcomes this problem: here the residual distribution is an exponential function of both the average residual variance and  $A_v$ , thus ensuring positive values [35]. Yet another option is to envision the microenvironmental sensitivity to be a special case of the reaction norm model, assuming an unknown environmental  $x$ -variable.

Selection for decreased heterogeneity is of interest for traits where large variation is undesirable. This might be true for, e.g., carcass traits where uniformity is favorable for the slaughtering process, but in general for traits with an intermediate optimum. When the mean performance is close to the optimum, more of the selection pressure will move toward reducing the residual variance [36]. Selecting for reduced residual variation might also be a way to select for more robust



animals, animals that can cope with unknown changes in the environment [33]. It would also be interesting to study the relationship between macro- and microenvironmental sensitivity.

### Consequences for Breeding Programs

The first obvious consequence of re-ranking  $G \times E$  (Fig. 2d) is of course that if you select individuals to become parents based on information from one environment and plan to use them in another environment, you have partially selected the wrong animals. One example of this could be a nucleus herd with very good environment but where the production animals are used under less optimal commercial conditions. The equally obvious solution is to include information from the production environment into the genetic evaluation. Mulder and Bijma [37] showed that the loss in genetic gain (compared with if there was no  $G \times E$ ) was lowest when progenies of males were tested in the production environment, and that this system was better than testing of sibs in that environment, and both systems were superior to testing only in the selection environment.

Another question is whether two breeding programs should cooperate even if their breeding goals differ to a certain extent. This could apply to two countries or to different production systems, e.g., organic and conventional production. It has been shown that long-term cooperation was beneficial if the correlation between breeding goals were higher than 0.8–0.9. However, initial cooperation was beneficial even when the correlation was as low as 0.4–0.6, but after some generations, the two populations had diverged so much that selection was practiced only within population. Furthermore, small breeding programs benefited from long-term cooperation at values of genetic correlation also below 0.8 [38]. It was also shown that for dairy cattle breeding programs, the genetic correlation had to be lower than 0.6 before it was beneficial to run two separate breeding programs rather than progeny testing bulls in both environments. If the selection intensity was high, the genetic correlation increased to 0.7–0.8, below which separate programs were optimal. Again, for a small population (e.g., organic or niche production), the genetic correlation had to be even lower before two programs were optimal [39].

So, in summary, it seems that even if there is  $G \times E$ , it is beneficial to cooperate and use information from other environments, unless the  $G \times E$  is extremely large. This is not the same as ignoring the existence of  $G \times E$ , rather it is to acknowledge its existence as something natural, and then adapting to that reality.

### Future Directions

The trend toward globalization of breeding will emphasize the importance of  $G \times E$ . A breeding company that wants to be successful in a global setting must also take into account what environment the animal is going to encounter. This might be even more problematic with the advent of genomic selection because it requires reference populations with phenotypic observations from the appropriate environments, if  $G \times E$  exists. And, in a global setting, there is no doubt that  $G \times E$  exists to a large extent, e.g., between the industrialized production systems in Europe or USA and systems in tropical environments in Africa. Because an increase in production efficiency in developing countries is both a key factor in poverty alleviation and to decrease the ecological footprint, it is important that genetic improvement is aimed at the appropriate traits expressed in the appropriate environments.

### Bibliography

#### Primary Literature

1. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman Group, Essex, UK
2. Pigliucci M (2001) Phenotypic plasticity: beyond nature and nurture. The Johns Hopkins University Press, Baltimore
3. Falconer DS (1952) The problem of environment and selection. *Am Nat* 86:293–298
4. Dawkins R (1982) The extended phenotype: the long reach of the gene. Oxford University Press, Oxford
5. Schlichting C, Pigliucci M (1998) Phenotypic evolution: a reaction norm perspective. Sinauer, Sunderland
6. Waddington CH (1942) Canalization of development and the inheritance of acquired characters. *Nature* 150:563–565
7. de Jong G, Bijma P (2002) Selection and phenotypic plasticity in evolutionary biology and animal breeding. *Livest Prod Sci* 78:195–214
8. Scheiner S (1993) Genetics and evolution of phenotypic plasticity. *Annu Rev Ecol Syst* 24:35–68
9. Scheiner SM, Lyman RF (1991) The genetics of phenotypic plasticity II. Response to selection. *J Evol Biol* 4:23–50
10. Scheiner SM, Lyman RF (1989) The genetics of phenotypic plasticity I. Heritability. *J Evol Biol* 2:95–107

11. Robertson A (1959) The sampling variance of the genetic correlation coefficient. *Biometrics* 15:469–485
12. Kirkpatrick M, Heckman N (1989) A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *J Math Biol* 27:429–450
13. Kirkpatrick M, Lofsvold D (1989) The evolution of growth trajectories and other complex quantitative characters. *Genome* 31:778–783
14. Gilmour A, Thompson R (2006) Statistical models for multidimensional (longitudinal/spatial) data. In: *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, Belo Horizonte, Brazil
15. de Jong G (1995) Phenotypic plasticity as a product of selection in a variable environment. *Am Nat* 145:493–512
16. Kolmodin R (2003) Reaction norms for the study of genotype by environment interaction in animal breeding. Swedish University of Agricultural Sciences, Uppsala
17. Strandberg E et al (2000) Genotype by environment interaction in Nordic dairy cattle studied by use of reaction norms. *Interbull Bull* 25:41–45
18. Kolmodin R et al (2002) Genotype by environment interaction in Nordic dairy cattle studied by use of reaction norms. *Acta Agric Scand A Anim* 52:11–24
19. Calus M, Groen AF, De Jong G (2002) Genotype  $\times$  environment interaction for protein yield in Dutch dairy cattle as quantified by different models. *J Dairy Sci* 85:3115–3123
20. Calus M, Bijma P, Veerkamp RF (2004) Effects of data structure on the estimation of covariance functions to describe genotype by environment interaction in a reaction norm model. *Genet Sel Evol* 36:489–507
21. Su G et al (2006) Bayesian analysis of the linear reaction norm model with unknown covariates. *J Anim Sci* 84:1651–1657
22. Kolmodin R et al (2004) Reaction norms for protein yield and days open in Swedish Red and White dairy cattle in relation to various environmental variables. *Acta Agric Scand A Anim* 54:139–151
23. Fikse WF, Rekaya R, Weigel K (2003) Genotype by environment interaction for milk production in Guernsey cattle. *J Dairy Sci* 86:1821–1827
24. Zwald N et al (2003) Identification of factors that cause genotype by environment interaction between herds of Holstein cattle in seventeen countries. *J Dairy Sci* 86:1009–1018
25. Raffrenato E et al (2003) Genotype by environment interaction for yield and somatic cell score with alternative environmental definitions. *J Dairy Sci* 86:2470–2479
26. Ravagnolo O, Misztal I (2000) Genetic component of heat stress in dairy cattle, parameter estimation. *J Dairy Sci* 83:2126–2130
27. Windig JJ, Calus M, Veerkamp RF (2006) Influence of herd environment on health and fertility and their relationship with milk production. *J Dairy Sci* 88:335–347
28. Meyer K (2001) Estimating genetic covariance functions assuming a parametric correlation structure for environmental effects. *Genet Sel Evol* 33:557–585
29. Calus M, Veerkamp RF (2003) Estimation of environmental sensitivity of genetic merit for milk production traits using a random regression model. *J Dairy Sci* 86:3756–3764
30. Pollott GE, Greeff JC (2004) Genotype  $\times$  environment interactions and genetic parameters for fecal egg count and production traits of Merino sheep. *J Anim Sci* 82:2840–2851
31. Fikse WF, Rekaya R, Weigel K (2003) Assessment of environmental descriptors for studying genotype by environment interaction. *Livest Prod Sci* 82:223–231
32. Namkoong G (1985) The influence of composite traits on genotype by environment relations. *Theor Appl Genet* 70:315–317
33. Mulder HA (2007) Methods to optimize livestock breeding programs with genotype by environment interaction and genetic heterogeneity of environmental variance. In: *Animal Breeding and Genetics Group*, Wageningen University, Wageningen
34. Mulder HA, Bijma P, Hill W (2007) Prediction of breeding values and selection responses with genetic heterogeneity of environmental variance. *Genetics* 175:1895–1910
35. SanCristobal-Gaudy M et al (1998) Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. *Genet Sel Evol* 30:423–451
36. Mulder HA, Bijma P, Hill W (2008) Selection for uniformity in livestock by exploiting genetic heterogeneity of residual variance. *Genet Sel Evol* 40:37–59
37. Mulder HA, Bijma P (2005) Effects of genotype by environment interaction on genetic gain in breeding programs. *J Anim Sci* 83:49–61
38. Mulder HA, Bijma P (2006) Benefits of cooperation between breeding programs in the presence of genotype by environment interaction. *J Dairy Sci* 89:1727–1739
39. Mulder HA et al (2006) Optimization of dairy cattle breeding programs for different environments with genotype by environment interaction. *J Dairy Sci* 89:1740–1752

## Books and Reviews

- Annicchiarico P (2002) Genotype  $\times$  environment interaction: challenges and opportunities for plant breeding and cultivar recommendations. *FAO Plant Production and Protection Paper* (174). Food and Agriculture Organization of the United Nations, Rome
- Hoffmann I (2010) Climate change and the characterization, breeding and conservation of animal genetic resources. *Anim Genet* 41:32–46
- Pigliucci M (2001) Phenotypic plasticity: beyond nature and nurture. The Johns Hopkins University Press, Baltimore
- Schlichting C, Pigliucci M (1998) Phenotypic evolution: a reaction norm perspective. Sinauer, Sunderland
- Thiruvankadan AK, Panneerselvam S, Prabakaran R (2010) Layer breeding strategies: an overview. *Worlds Poult Sci J* 66(3):477–501

## Animal Molecular Genetics from Major Genes to Genomics

ASKO MÄKI-TANILA

Biotechnology and Food Research, MTT Agrifood Research Finland, Jokioinen, Finland

### Article Outline

Glossary

Definition of the Subject

Introduction: From Infinitesimal Model to Major Genes

QTL Mapping

Genes Underlying QTL

Genomic Revolution

Future Directions

Acknowledgment

Bibliography

### Glossary

**Additive genetic variance** is usually the largest part of genetic variation in the quantitative trait and is due to average effects of genes. The change in mean caused by selection is proportional to the additive genetic variance.

**BLUP** is a short name for Best Linear Unbiased Prediction. It is now a norm in estimating breeding values within populations. It uses information on all kind of relatives and corrects the data for differences in production environment.

**Breeding value** of an individual is the expected value of its progeny relative to the population mean.

**Candidate gene** is a possible mutation underlying the mapped QTL. A positional candidate is a gene located in the same region as a mapped QTL. A biological candidate for a QTL is a gene which has a function related to the quantitative trait.

**Effective population size** ( $N_e$ ) is the number of individuals that with random mating result in the same rate of inbreeding as the population itself.

**Genetic marker** is a specific detectable sequence of DNA with a known location in the genome.

**Heterosis** Is the extent to which the performance of crossbred animals is better than the mean of two parental populations.

**Infinitesimal model** assumes the genetic variation of a quantitative trait is due to infinitely many unlinked genes each with an infinitesimally small additive effect, so that selection produces negligible changes in gene frequency and variance at each locus.

**Linkage disequilibrium** is a non-random association of alleles across loci. Recombination between loci will gradually reduce the associations, more slowly the closer the loci are to each other. Adjacent markers with correlated allele frequencies could be used for mapping and selection.

**Marker-assisted selection** is selection on a quantitative trait where also the information on associated markers is used as a selection criterion. Gene-assisted selection is a special case where the marker is at the major gene causing the variation.

**Mixed model equations** are providing a method to simultaneously solve the predicted breeding values (random effects) for animals and estimate the fixed effects due to differences in production environment.

**Morgan** is the unit for a map distance in the genome.

**Quantitative trait locus (QTL)** A short genomic region with a large effect on a quantitative trait.

**Single nucleotide polymorphism (SNP)** is variation caused by a mutation at a single nucleotide.

### Definition of the Subject

Animal breeding is a major contributor to the vast improvements in animal production over decades. The main tool in breeding operations is selection; now, more and more attention is also paid to the amount and nature of genetic variation. It is on these topics that the chapter is built on. It starts from the fundamental methods in determining the genetic value of animals. The normal distribution and linear methods stemming from the concept of large number of loci with tiny effects causing variation are the base line. For quite some time, there have been observations on major loci causing deviation in linear prediction of genetic values. There is a part introducing methods to get around such cases and turning them advantageous by directly utilizing the variation mediated by major

loci. The longest section is on introducing molecular genetic tools to find areas in the genome harboring such major loci and to further characterize the actual underlying genes. The most recent development in DNA technology is the availability of high throughput analysis of individual genomes with tens of thousands of signposts or markers. Without knowing the causal factors, the extensive marker panels can dissect the variation for named markers and recompose the piecewise information for prediction and selection for individuals with respective marker information. Thus, with a sufficiently large number of genotyped and phenotyped animals, selection decisions can be made immediately after birth and very speedy genetic progress is achieved. The same marker panels could be used to have a detailed picture on the state of genetic variation in a population and genome areas causing heterosis in crosses between populations. The theory of animal breeding is thoroughly introduced and discussed in many books, for example those by Falconer and Mackay [1] and Lynch and Walsh [22]. The recent developments in the utilization of DNA markers in characterizing the variation is well covered in the textbook by Weller [2].

### Introduction: From Infinitesimal Model to Major Genes

Animal breeding has been very successful in improving production efficiency through utilizing genetic variation between animals. Most of the economic traits are measurable and require own type of approach in analyzing the genetic variation, that is, quantitative genetics. Quantitative genetics theory provides a statistical description of the genetic and environmental variation affecting a particular measurement in a random breeding population as it is at the moment and allows some short-term predictions of the response to selection. In essence, the *additive genetic variance* and its related concept, heritability ( $h^2$ ), give a coherent framework into which observations on individuals and population, such as the effect of selection or the similarity between relatives, can be fitted.

When there is additive gene action across loci and random mating, the contributions from different loci will be independent, so that their sum will become asymptotically normal as the number of loci increases

(while each locus has an infinitesimally small effect). This is referred to as an *infinitesimal model*. Also, the joint distribution of parent and offspring – or of any related individuals – will be approximately normal when the number of loci is large. Therefore, linear methods could be used for prediction. Animal breeding data is typically influenced by many nongenetic factors. The Best Linear Unbiased Prediction (BLUP), due to Henderson [3], has been developed to simultaneously handle both fixed (environmental) and random (mainly genetic) effects in a mixed model. In symbols, denoting individual performance record by  $y$ , production environment effects by  $\beta$  and individual animal's genetic value by  $u$ , there is an equation for the genetic value of the  $i$ th animal (with known sire and dam and  $n$  progeny of varying genetic value  $u_p$  with known mates)

$$\beta_i + (1 + 2k + \frac{1}{2}kn)u_{ij} - k(u_{\text{sire}} + u_{\text{dam}}) + \sum_{\text{progeny}} (\frac{1}{2}u_{\text{mate}} - u_p)k = y_{ij} \quad (1)$$

where  $k = (1 - h^2)/h^2$ . The conventional *breeding value* estimation model for a large number of animals would contain the observation vector  $\mathbf{y}$  explained by  $\mathbf{u}$  of random additive genetic effects and  $\mathbf{e}$  of random residual effects simultaneously correcting for the fixed environmental effects  $\boldsymbol{\beta}$ . The fixed effects and random genetic effects are connected to observations with the incidence matrices  $\mathbf{X}$  and  $\mathbf{Z}$ , respectively, and  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ . The genetic values are made of additive effects within and between loci, dominance deviations due to interaction within loci, and epistatic deviations due to interaction between loci. The covariance matrix of additive genetic effects can be expressed with the additive relationship matrix ( $\mathbf{A}$ ) as  $\mathbf{A}\sigma_a^2$ . For solving the *mixed model equations*, one needs  $\mathbf{A}^{-1}$  which is actually easier to form than the matrix itself [4]. From the inverse elements, one can extract equations like (1). Indeed, an approximate BLUP based on the consideration of a chosen set of close relatives gives satisfactory accuracy in comparing different selection schemes [5].

Matings between relatives cause inbreeding (increased probability for homozygosity) and inbreeding depression in loci with segregating recessive alleles with undesirable effects. In order to improve the genetic evaluation, the *mixed model equations* should also include dominance effects, although the inverse of

the joint covariance matrix of additive and dominance effects is very challenging under inbreeding [6]. It is indeed the inbred populations where the influence of dominance is more pronounced, or the crosses of inbred animals recovering from inbreeding depression or showing *heterosis*.

### Selection at the Locus Level

All the concepts of quantitative variation can be expressed at the gene level [1], and therefore it is possible to describe how selection is altering allele frequencies of any locus causing variation in the quantitative trait. The genetic variance is made of contribution from individual loci. The change in mean or response to selection is proportional to the *additive genetic variance*. Hence, the contribution of any locus (or several loci) to the genetic progress under selection will be in proportion to its (their) contribution to the overall *additive genetic variance*. If a locus contributes 10% of the genetic variance, it will explain the same 10% of the total response to selection. Modifying the illustration in Alan Robertson's unpublished lecture notes at the University of Edinburgh, assume that in a dairy cattle population, there is an additively acting biallelic locus where the difference in average milk yield between two homozygotes is  $2a = 400$  kg. Further suppose that the two alleles have the same allele frequency of 0.5. If the selection lifts the mean milk yield by 500 kg, how much has the frequency ( $x$ ) of the allele with increasing effect changed at the locus. Before selection, the locus contributes to the *additive genetic variance* by  $\frac{1}{2} x (1 - x) (2a)^2 = 20,000$ . The total genetic variance is – say – 250,000, so that the current locus controls 8% of the total variance and will therefore be responsible for a selection change of 40 kg. This kind of change could be caused by a change 0.1 in allele frequency so that after selection the allele frequency in the progeny generation is 0.6.

The relationship between individual loci and selection response could be described yet another way. The population must respond to selection with an increase in the frequency of those alleles which increase the trait value. The population genetic theory says that the change in gene frequency caused by selection is  $\frac{1}{2} s x (1 - x)$  where  $s$  is the selective advantage of one homozygote over the other. The aim is to find in the

quantitative genetic context the selective advantage of the desirable homozygote compared to the other homozygote after imposing a certain selection intensity ( $i$ ) for a measured trait. Or to know what is the relationship between a genetic effect on the quantitative trait and the consequent selective advantage caused by the selection on the trait. Falconer and Mackay [1] are showing that there is a linear relationship between the two so that  $s = i 2a/\sigma_p$  where  $\sigma_p$  is the phenotypic standard deviation, so the change in the population mean is equal to  $2a$  times the change in allele frequency  $= 2 i a^2 x (1 - x)/\sigma_p$ . Since  $2 a^2 x (1 - x)$  is the *additive genetic variance* due to the locus, the response produced by a change in allele frequency at the locus is proportional to the *additive genetic variance* caused by the locus. Summing over all loci to compute the total response yields again the classical formula for the selection response  $\Delta G = i \sigma_a^2 / \sigma_p$ .

Regarding the example and writing  $\sigma_p = 1000$  (assuming  $h^2 = 0.25$ ), if the selection response of 500 kg had taken place in just one generation, it would have to be caused by an intensity  $i = 2$ . The expected change of frequency at the studied locus then equals  $\frac{1}{2} i 2ax (1 - x)/\sigma_p = 0.1$ . The example has been on a simple additive locus. The same exercise could be done for dominance – or even for epistasis – considering the average effect of allele substitution on the mean of individuals within the population and the effect of small changes of allele frequency on the population mean.

### Major Genes

Genes with smaller effects may have their gene frequencies altered less by selection. Highly selected populations might therefore be expected to be segregating for loci with smaller effects on a selected trait compared to those with no previous history of selection. For loci with a tiny effect on the characteristic, directional changes of gene frequency due to selection will be low and changes will occur mostly by chance, depending on the number of parents used each generation.

Loci with very large effects on the trait will have their frequencies changed much more by selection. Since their contribution to the genetic variance depends on the allele frequency, the genetic variance will also change. The immediate effect will depend on



the starting frequencies in the population. When the allele effects are large, it is necessary to include also second-order terms of  $a/\sigma_p$  for the allele frequency change (e.g., [7]) and the respective response would have a quadratic terms of phenotypic values (or selection intensity). These effects would be more pronounced when there are dominance effects between the alleles. In case of carrying selection in one sex only, for example, if the variance is due to a locus with complete dominance and  $x = 0.9$  and  $a/\sigma_p = .5$ , the response upward is going to be overestimated with linear prediction by up to 31% and 21%, when the proportion selected is 10% and 20%, respectively [8]. Also the *mixed model equations* would work less efficiently when the variation is caused by a small number of loci [9].

Over the decades, there have been findings on major genes segregating in farm animal populations: the halothane gene mutation causing stress syndrome and leanness in pigs [10], dwarf affecting body size in chicken [11], double muscling in beef cattle [12], booroola lifting prolificacy in sheep [13]. There is clear evidence that the contributions from the loci affecting the quantitative variation are vastly unequal, there being a small number of major loci and a larger number of minor loci. Such an understanding has generated research on developing tools to utilize major genes and to detect them with new available mapping tools made available by molecular genetics.

### Improvement Schemes Utilizing Major Genes

The previous discussion shows that segregating major genes may bias the linear prediction of response in selection. On the other hand, a direct selection on a gene with large effect should give higher overall gains in a selection scheme. If one is able to type the different genotypes at the major locus, it is possible to carry out gene-assisted selection. Until recently, this has not been possible and selection has been based on marker loci close to the actual gene, therefore the term “*marker-assisted selection*” (MAS). Earlier the markers were typed by utilizing protein variants, now the markers are solely based on DNA polymorphism.

While conventional selection is favoring desirable alleles in an indirect way, it should be beneficial to have a tool which can influence the allele frequencies in

a much more direct way. The ordinary selection with intensity  $i$  would change the mean by  $i h \sigma_a$ . Denoting the variance due to the major locus with  $\sigma_M^2$  and its size relative to the total *additive genetic variance* as  $R$ , it is possible to assess the advantages of exploiting the marker information in selection.

Lande and Thompson [14] used the selection index methodology in investigating the influence of gene or marker-assisted selection. The selection can be based both on phenotype ( $P$ ) and marker genotype ( $M$ ). First step is to compute the selection weights  $b_1$  and  $b_2$  in the index

$$b_1 \text{ phenotype} + b_2 \text{ marker genotype} = b_1 P + b_2 M$$

This could be used to generate equations for the covariance of the index with  $P$  and  $M$  (the genotypic value of the trait is denoted with  $G$ ).

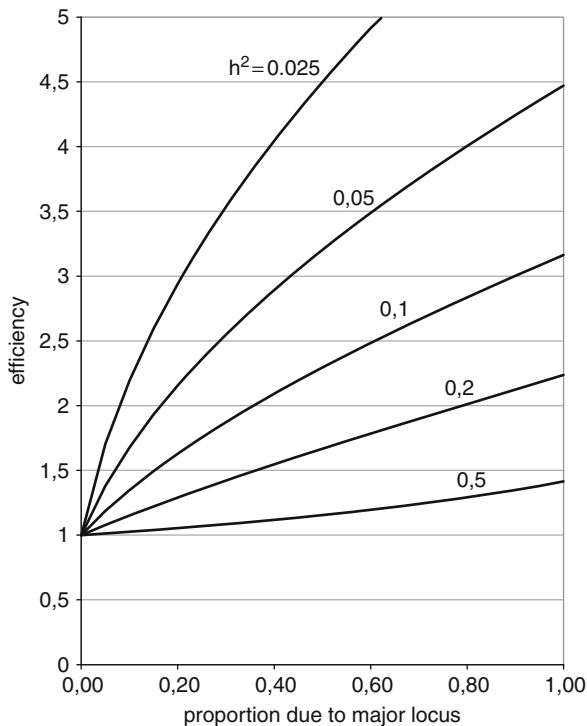
$$b_1 \text{ cov}(P, P) + b_2 \text{ cov}(M, P) = \text{cov}(G, P)$$

$$b_1 \text{ cov}(P, M) + b_2 \text{ cov}(M, M) = \text{cov}(G, M)$$

where  $\text{cov}(P, P) = \sigma_p^2$ ,  $\text{cov}(G, P) = \sigma_a^2$ ,  $\text{cov}(P, M) = \text{cov}(G, M) = \text{cov}(M, M) = \sigma_M^2$ . The weights in the selection index are  $b_1 = h^2(1 - R)/(1 - Rh^2)$  and  $b_2 = (1 - h^2)/(1 - Rh^2)$ . The correlation between the index and genotypic value is approximately  $h \sqrt{1 + R/h^2}$ . In conclusion, the marker information would be useful, when  $R$  is high or when a major part of the variation is due the QTL and when the heritability of the trait is low (Fig. 1). Therefore, the selection of low heritability traits such as longevity traits (fertility or disease resistance) would benefit if useful genes or linked markers could be detected.

There are traits which are expressed only in one sex. The use of markers would be useful for such traits. Dairy cattle breeding is a good example: markers would allow pre-selection of bulls on milk traits which would mean savings in progeny testing. Selection could be carried out as early as the embryo stage, resulting even higher reductions in generation interval.

Genes or markers are also beneficial for traits that are difficult to improve under traditional selection. Good examples on such are carcass traits that require slaughtering before they can be measured. Meat pH, tenderness, and color are typical carcass traits. Traits that are available late in life, like longevity and lifetime fecundity, would benefit from MAS. The measurement



**Animal Molecular Genetics from Major Genes to Genomics. Figure 1**

Efficiency of marker-assisted selection relative to conventional phenotypic selection with the same selection intensity in large population with respect to the proportion of variance explained by a major gene, for different heritability values of the trait

and data collection of some traits, for example, disease resistance, is very expensive and therefore there is much interest in finding major genes for them.

In general, MAS has advantage over conventional selection for alleles that are initially rare and for alleles that are recessive. It is customary to think that the variation in a quantitative trait is caused by a large number of nongenetic factors. However, the marker genotype would be accurately known without environmental noise.

The consideration of gene or *marker-assisted selection* could therefore include the following points: difficulties of collecting performance records, heritability of the trait, proportion of variance explained by the major gene, and the availability of performance in the life cycle. Methods and applications of MAS are comprehensively reviewed by Dekkers [15].

The *marker-assisted selection* has also drawbacks. The direct selection on a gene would be efficient only for a short time. The desirable allele would quickly become common and new beneficial alleles should be found or new potential markers should be searched for useful segregation. If the selection is not directly on the gene, it is possible that the recombination between the used marker and QTL would gradually weaken the association. The use of markers would even be harmful if the association is not properly validated and efficiency of selection reduced compared to conventional selection.

There are also consequences for the overall genetic variation. The larger the effect of the QTL is, the faster it is fixed by selection and the more it will also reduce the variation in the surrounding genome area (e.g., [16]). An accelerated increase in the marker allele frequency may be accompanied by very unbalanced use of family lines. This is increasing the risks for the reduction of variance and enrichment of harmful recessives.

### Introgression from Other Populations

Many local breeds carry interesting genetic variants that are considered beneficial to be introduced to a commercial main stream breed. The former is termed as a donor breed and the latter one as a recipient breed. The operation is usually done by an introgression program. It consists of forming an initial cross between the breeds followed by repeated backcrosses to the recipient one to recover the economically important genome. The target gene is maintained in the backcross generation through selection of donor gene carriers. After some generation of backcrossing the program will finish by a generation of intercrossing to make the population homozygous for the desired allele.

Genetic markers could be useful in introgression programs in two ways [17]. First, markers can be used to select individuals at each backcross generation which are heterozygous for the desired allele or homozygous in the last generation of intercross (foreground selection). Secondly, markers can be used to enhance the recovery of the recipient genome (background selection).

The introgression strategy is a popular genetic improvement method in plants, while there are still

very few applications in domestic animals. An example is the naked neck gene in chicken which reduces plumage in chicken and makes animals more tolerant to heat. It was introgressed from low body weight landrace chickens into a commercial meat-type Cornish chicken [18].

Most of the economic traits in farm animals are complex ones with polygenic variation. One of the most promising cases of introgressing exotic germplasm to commercial production populations has been the impact made on western pig production by the Taihu breeds of the Shanghai area (China). The background for the interest and the results have been reviewed by Knap and Neeteson [19]. Some of these pig breeds (Meishan and Jiaying) were imported into France, the UK, and the USA in the 1980s. The interest is in reproductive traits: low age at puberty, high ovulation rate and embryo survival resulting in high litter size, and high teat numbers. The disadvantage is in high body fat levels and slow growth rates.

Several commercial breeding organizations in France, UK and Canada have invested in the introgression of Jiaying and particularly Meishan pigs into their dam lines. This was accompanied by studies into the trade-off between improved fertility and reduced leanness (e.g., [20]). These suggested that the feasible way to commercially exploit these genotypes would be to have commercial sows with 12.5% Taihu genes [21].

While the fertility is improved, as a compromise, leanness is reduced in Taihu-based genotypes, which is a serious obstacle toward large-scale commercial exploitation, although persistent genetic improvement in the latter trait obviously would solve the problem. A complicating factor here is that a few years after the Taihu imports, large data bodies and improved statistical methods allowed for a faster genetic change of reproductive traits in pigs, while genetic improvement of leanness continued at the same rate as before.

Western commercialization of the Taihu breeds has been successful so that almost 100,000 Taihu-based parent gilts are currently sold per year in Europe and North America. It represents less than 1.5% of the total market volume. A good working example on successful introgression of exotic genotypes into advanced animal breeding program is still missing for a polygenic trait.

## QTL Mapping

### Markers and Linkage Maps

Better understanding and utilization of major genes would benefit from the location of genes or genome areas responsible for the variation in quantitative traits. The prerequisite for this is to have a reasonable skeleton of the genome sites to survey and locate the findings. The genome is made of chromosomes. There are two homologues of each chromosome in a typically diploid animal genome. One homologue is originating from sire (paternal chromosome) and the other one from dam (maternal chromosome). The alleles at the loci on the paternal (maternal) chromosome tend to be inherited together or they are said to be linked. The genome is full of variable sites which could be used as markers to construct a detailed map of locations for further work. The closer the markers are to each other, the more likely they are inherited together or the tighter is the linkage between them or the rarer are the recombinants between them. The distant markers are on the other hand showing independent segregation. Physically, recombination is seen as a crossing over of chromosome strands. The chromosomes could be termed as linkage groups.

The linkage is measured as recombination frequency ( $c$ ) or the proportion of recombinant ones among all the gametes. If markers are very far from each other, they may be several crossing-overs between them. When the loci are freely combining, the number of recombinant and nonrecombinant gametes is equal and  $c = 0.5$ . The map distance is expressed in *Morgans* (M) with its hundredth being cM (centimorgan). A low recombination frequency 1% corresponds to 1 cM. For longer distances, the possibility of several crossing-overs should be considered, as, for example, two crossing-overs between distant markers would often result in nonrecombinants of the marker pair itself.

The detection of linkage or computing map distances would require variable sites and use of heterozygotes. Ideal markers for mapping are the (codominant) loci with alleles not showing dominance (or recessivity). Linkage could be demonstrated with different mating options. For example, there can be double heterozygote  $M_1 N_1 / M_2 N_2$  mated with a double homozygote  $M_1 N_1 / M_1 N_1$ , in which case

the offspring numbers in different genotype classes would give a straightforward measure for linkage. Inbred lines are homozygous for all loci and the first generation cross  $F_1$  heterozygous, respectively. When  $F_1$  animals are crossed with one of the parental lines, the resulting backcross animals would give direct estimates on recombination frequencies. Using inbred line terminology, the mating is called double backcross.

Although there are often heterozygotes available, only a limited number of cases are informative on possible recombinants and determining the amount of linkage. There are five different cases with respective information content.

|                                   |           |                                   |                                 |
|-----------------------------------|-----------|-----------------------------------|---------------------------------|
| $\frac{M_1}{M_1} \frac{N_1}{N_1}$ | $\otimes$ | $\frac{M_1}{M_1} \frac{N_1}{N_1}$ | Not informative                 |
| $\frac{M_1}{M_2} \frac{N_1}{N_1}$ | $\otimes$ | $\frac{M_1}{M_1} \frac{N_1}{N_1}$ | Not informative                 |
| $\frac{M_1}{M_2} \frac{N_1}{N_2}$ | $\otimes$ | $\frac{M_1}{M_1} \frac{N_1}{N_1}$ | Informative (double backcross)  |
| $\frac{M_1}{M_2} \frac{N_1}{N_2}$ | $\otimes$ | $\frac{M_1}{M_2} \frac{N_1}{N_1}$ | Informative (single intercross) |
| $\frac{M_1}{M_2} \frac{N_1}{N_2}$ | $\otimes$ | $\frac{M_1}{M_2} \frac{N_1}{N_2}$ | Informative (double intercross) |

From the last mating type, there are nine different progeny classes with respective expected numbers according to the recombination frequency.

Before a marker is used, its suitability is assessed studying the allele frequencies ( $x_i$ ). So far it is obvious that homozygous parents are of no use and among the progeny of heterozygous parents only the homozygotes (or for multiple alleles, the heterozygotes  $M_1/\text{no } M_2$  or  $M_2/\text{no } M_1$ ) would signal which allele of the parent has been transmitted to the progeny. So the quality of the marker is based on polymorphism information content (PIC), which is the probability of identifying which homologue of a given parent was transmitted to a given offspring, the other parent being genotyped as well, or in other words, probability that the parent is heterozygous  $\times$  probability that the offspring is informative. Summing over alleles, the PIC for a multiallelic marker would be  $1 - \sum x_i^2 - 2 \sum \sum_{i < j} x_i^2 x_j^2$ .

### Mapping Function

In order to fill the linkage map with markers and genes, there is a need to have a good measure for the map distance. As there can be several crossing-overs between distant markers, recombination frequency as such is

not an appropriate measure. If there are markers  $M$ ,  $N$ , and  $O$  (with respective recombination frequencies  $c_{MN}$ ,  $c_{MO}$  and  $c_{NO}$ ) and crossing-overs in the adjacent genome regions would be independent, then the recombination frequency  $c_{MO} = c_{MN} + c_{NO} - 2 c_{MN} c_{NO}$ . In other words, the recombination frequencies are additive only if they are small enough so that the product term could be ignored. A map distance is required that would give the total number of crossovers between even very distant markers. The measure should be additive so that the number of crossovers between  $M$  and  $O$  is the number of crossovers in the interval  $M-N$  plus those in the interval  $N-O$ .

The probability of no crossovers is  $e^{-\lambda}$  (from the Poisson distribution) and the probability of at least one recombinant gamete is  $c = \frac{1}{2}(1 - e^{-\lambda})$ . The latter is true, because for each crossover event only one-half of the gametes will be recombinant types there being only two of four strands involved in a crossover. The map distance is then  $-\frac{1}{2} \ln(1 - 2c)$  in *Morgans* for an observed recombination rate  $c$ . For short chromosome segments ( $c < 10\%$ ), the map distance equals the recombination rate, while 50 cM corresponds to recombination rate 32%. The approach is called Haldane mapping function and is valid when the crossing-overs across the genome occur independently. If some degree of interference needs consideration, Kosambi mapping function should be applied. For more details, see Lynch and Walsh [22].

The number of chromosomes varies quite much over the animal species:

| Cattle | Chicken | Pig | Sheep | Carp | Human | Mouse |
|--------|---------|-----|-------|------|-------|-------|
| 30     | 39      | 17  | 27    | 52   | 23    | 32    |

When there are so many chromosomes in livestock species, the average recombination across the genome is very high. In comparison, the well-known lab organism *Drosophila* has only four chromosomes and no recombination in males.

The genome in mammals is about  $3 \times 10^9$  b (base) and 3 M. The chicken genome is only about  $1 \times 10^9$  b, while the linkage map is still around 3 M. The well-known lab organism *Drosophila* has only 275 cM and  $0.13 \times 10^9$  b.

## DNA Markers

Ideal markers for mapping purpose should be widely distributed over the genome. At the same time, it is important that there are several markers available within short genomic regions allowing high-resolution mapping. Good markers should have many alleles, and with frequencies as equal as possible. The alleles should be codominant to allow easy detection of heterozygotes. The allele typing should be straightforward with satisfactory repeatability across analyses. Alleles should also be stable from generation to generation with a low mutation rate. The cost of genotyping could be reduced if the genotyping could be done simultaneously for many loci and the analysis could be easily automatized to provide high throughput.

The most common markers are either microsatellites (based on loci which have a variable number of repeats of a same short sequence) or single nucleotide polymorphism (SNP, which is a nucleotide site in the genome sequence, predominantly biallelic, but possibly having four alleles with each of the four nucleotide bases appearing in the same location). Both these marker types are widely distributed and codominant. Microsatellites are now less popular because their repeatability is lower and there is a limited number of them for a high-resolution mapping in a chosen short region. SNPs have gained popularity and it is nowadays possible to have very large SNP panels typed moderately cheaply in a single analysis with a small microarray chip.

Microsatellites are still used when sufficient information is delivered by only few markers (e.g., as required for parentage testing), and in some applications, this will offset the relative disadvantage of throughput in comparison to SNP markers.

## From Segregation Analysis to Use of Markers

Since the genes affecting quantitative genetic variation are behaving like genes with classical discrete genotypes, there have been attempts to characterize them. If in the population there is segregation at a major locus mediating quantitative trait variation, alleles with very large effect and extreme frequency could be causing a skewed phenotypic distribution. Even if allele classes are equally common, the distribution may have several peaks. The consequences due to major gene can be

studied with segregation analysis (e.g., [23]). Segregation analysis is full of factors which are hard to be itemized and tested as the effect of several major genes, extreme allele frequencies and dominance gene action produce similar distribution patterns and are therefore hard to separate. Segregation analysis is historically important and would still be used for providing preliminary understanding about the inheritance pattern, especially for human familial diseases.

If one wants to demonstrate the existence of a major gene, it is better to anchor the analysis of distribution features to a known locus with visible genotype classes, and hope that the locus is a useful marker to demonstrate the existence of a linked major gene. Thereby, the analysis could be narrowed down to investigating one gene (or marker) at the time (Fig. 2). The statistical model could include the location of the gene, the effect size, dominance, and frequency of the alleles. Only the availability of marker loci supporting the localization was for a long time restricting the possibility for such a work.

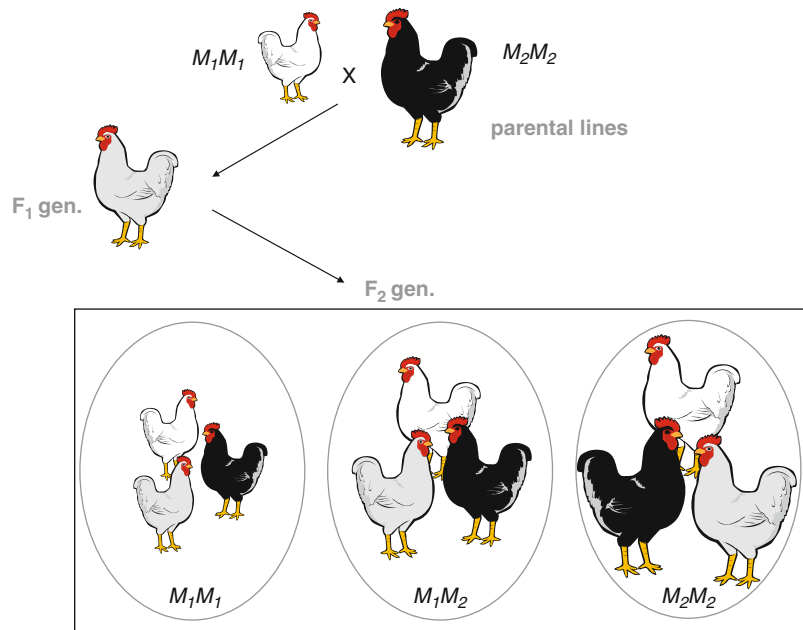
It was half a decade ago when the blood group alleles became immunologically identifiable and possible signposts for analyzing quantitative genetic variation in farm animals. These markers were followed by electrophoretically (molecule size and electric charge) detected protein. The number of markers was still very low and even their rough location was not known. Enthusiastic researchers were facing frustrating results in hunting major genes, although the possibilities and optimum designs for analyzing quantitative traits were clearly envisaged by some groups (e.g., [24]).

The research chances improved a lot when molecular genetic techniques opened completely new kind of possibilities for mapping. With a lot of markers available all around the genome, it was soon rather rewarding to detect regions responsible for the variation. Also a new term, QTL or *quantitative trait locus*, for such genomic regions was coined.

## QTL Mapping from Crosses of Inbred Lines

Inbred lines would be ideal resource populations for QTL mapping. They are homozygous for most of the genome, including, hopefully, the used marker locus and putative QTL, in which case the crossbred F<sub>1</sub> population would be heterozygous at both the loci. In other





**Animal Molecular Genetics from Major Genes to Genomics. Figure 2**

Two chicken lines differing in body weight and plumage color are fixed for a marker allele ( $M_1$  or  $M_2$ ). The lines are crossed to follow the possible co-segregation of the traits and also of the marker locus in the  $F_2$  animals where all possible combinations of phenotypic classes and marker alleles are appearing. The animals are grouped by the genotypes  $M_1M_1$ ,  $M_1M_2$ , and  $M_2M_2$ . The  $M_1M_1$  animals are the heaviest, while the  $M_2M_2$  ones are the lightest with the  $M_1M_2$  ones being intermediate. This would suggest that the marker is close to a QTL related to body weight. However, the plumage color which is also showing a simple pattern of inheritance, is segregating independently as the color types occur with equal frequencies in the different weight classes

words in the  $F_1$  generation, there is a very high correlation of allele frequencies across the loci, or using population genetic terms, the *linkage disequilibrium* between the marker and possible QTL is maximized. *Linkage disequilibrium* is reduced over generations as recombination is gradually breaking the associations. However, very little of *linkage disequilibrium* will be lost within short map distances in producing backcross and  $F_2$  generation. Starting from inbred lines, it is easy to follow the variation in the marker loci and the allele phases across marker and QTL would remain the same over further generations. While the use of line crosses is enhancing marker variation, it is also beneficial to choose lines that are phenotypically very divergent. Such a starting point would increase the chances of finding QTL.

Assume there are parent lines ( $P_1$  and  $P_2$ ) homozygous for marker locus (two alleles  $M_1$  and  $M_2$ ) and QTL ( $B$  and  $b$ ). The crossed generations are mixtures of

distributions with respect to the marker genotypes. The line differences could be associated with the marker classes using alternative crossing designs. Then the expected distribution and contrasts for the marker genotypes between the lines are

|                      | Expected distribution among marker genotype classes |               |               |                                                             |
|----------------------|-----------------------------------------------------|---------------|---------------|-------------------------------------------------------------|
|                      | $M_1M_1$                                            | $M_1M_2$      | $M_2M_2$      | Contrasted classes                                          |
| Backcross with $P_1$ | $\frac{1}{2}$                                       | $\frac{1}{2}$ |               | $M_1M_1 - M_1M_2$                                           |
| Backcross with $P_2$ |                                                     | $\frac{1}{2}$ | $\frac{1}{2}$ | $M_1M_2 - M_2M_2$                                           |
| $F_2$                | $\frac{1}{4}$                                       | $\frac{1}{2}$ | $\frac{1}{4}$ | $M_1M_1 - M_2M_2$<br>$M_1M_1 - M_1M_2$<br>$M_1M_2 - M_2M_2$ |

The efficiency of associating the computed phenotypic differences with the marker variation depends on

how closely the analyzed single marker is to a gene affecting the variation. The  $F_2$  generation (produced by intercrossing  $F_1$  animals) is better than the backcrosses (crossing  $F_1$  with one of the parent lines) as it is possible to estimate also the dominance effect via comparing the heterozygote class with the homozygote ones.

### Random Breeding Populations

Typically, farm animals are forming a random breeding population. Such a population has many limitations for successful QTL mapping. The available phenotypes are within a relatively narrow range of variation and would generate much smaller class differences than in crosses from diverse inbred lines. The markers would also show less systematic polymorphism than crossbred populations, since in a large random breeding population the *linkage disequilibrium* is expected to be very low over moderate map distances. In that situation, the only chance to find QTL is to have a detectable marker at the gene itself or its immediate neighborhood. Still, some families or parent individuals may be useful for mapping studies as by chance they have marker and QTL variation phased in a useful manner. Although there is no *linkage disequilibrium* at the population level, there may be partial disequilibrium within families depending on the recombination rate. Such a *linkage disequilibrium* could be exploited in QTL screening. Assume sires that are heterozygous for the marker ( $M_1M_2$ ) and have large progeny group of half-sibs with phenotype and marker genotype data. The data could be modeled for an analysis of variance around the average level  $\mu$  as performance record =  $\mu$  + marker genotype + residual.

Using the notation above, at the QTL locus the separation of two homozygotes is  $2a$ , the frequency of  $B$  allele in the population is  $x$  and the recombination rate between the marker and QTL is  $c$ . The analysis could be extended to include dominance by allowing the heterozygotes to deviate from the mean of the two homozygotes by  $da$  ( $d = 1$  dominant,  $d = 0$  additive,  $d = -1$  recessive gene action with the intermediate values indicating partial dominance or recessivity).

Therefore the frequency of different progeny types would be across heterozygous sires with different QTL allele configuration as follows:

|               | QTL genotype in progeny |                       |                  |
|---------------|-------------------------|-----------------------|------------------|
|               | $BB$                    | $Bb$                  | $bb$             |
|               | 0                       | $(1 + d)a$            | $2a$             |
| Sire genotype | $M_1$ allele from sire  |                       |                  |
| $M_1B/M_2B$   | $x$                     | $1 - x$               | 0                |
| $M_1b/M_2b$   | 0                       | $x$                   | $1 - x$          |
| $M_1B/M_2b$   | $(1 - c)x$              | $(1 - c)(1 - x) + cx$ | $c(1 - x)$       |
| $M_1b/M_2B$   | $cx$                    | $c(1 - x) + (1 - c)x$ | $(1 - c)(1 - x)$ |
|               | $M_2$ allele from sire  |                       |                  |
| $M_1B/M_2B$   | $x$                     | $1 - x$               | 0                |
| $M_1b/M_2b$   | 0                       | $x$                   | $1 - x$          |
| $M_1B/M_2b$   | $cx$                    | $c(1 - x) + (1 - c)x$ | $(1 - c)(1 - x)$ |
| $M_1b/M_2B$   | $(1 - c)x$              | $(1 - c)(1 - x) + cx$ | $c(1 - x)$       |

From these frequencies, one can compute for different progeny groups across the population:

| Sire genotype | The expected phenotypic difference between $M_1$ and $M_2$ progeny |
|---------------|--------------------------------------------------------------------|
| $M_1B/M_2B$   | 0                                                                  |
| $M_1b/M_2b$   | 0                                                                  |
| $M_1B/M_2b$   | $(1 - 2c)[1 + (1 - 2x)d]a$                                         |
| $M_1b/M_2B$   | $-(1 - 2c)[1 + (1 - 2x)d]a$                                        |

Hence, the difference between subgroups disappears when recombination between the marker and QTL is 0.5 (or when the marker is very far from QTL). In general, it is not possible to separate the effect and location, for example, QTL which is 25 cM away from the marker would have an estimate like with the marker with half of the effect locating exactly at the QTL position. On the other hand, the frequency and dominance action will also affect the estimation and obviously the detection of rare recessives is very hard.

When at the population level there is no linkage disequilibrium, the occurrence of informative haplotypes in progeny depends on the allele frequencies at

the marker and QTL. The heterozygous sires  $M_1M_2$  are investigated for the difference in performance between  $M_1/\text{no-}M_2$  and  $M_2/\text{no-}M_1$  progeny classes. In principle, the analysis should pool the results across sires to improve the power of the analysis. However, there are sires (like  $M_1 B/M_2 B$ ,  $M_1 b/M_2 b$ ) where the marker genotype contrast is zero. In some cases, the difference is positive ( $M_1 B/M_2 b$ ) and in other cases negative ( $M_1 b/M_2 B$ ). Therefore, the analysis is carried out within sires. Thereby, it is possible to include all the possible existing gametes (and alleles) with different combinations of marker and QTL alleles, and at the same time eliminate the spurious association caused by between sire variation [24, 25].

### Interval Mapping

When the QTL screening is done marker by marker over the genome, the available information is utilized in a suboptimal way, and, for example, the effect and location of QTL cannot be separated. On the other hand, when there are many markers, some of them may by chance give a statistically significant result at the chosen risk level. If the significance threshold is 5%, then five out of hundred analyzed map positions may give a “significant” result. This could be taken into account by having a more stringent rejection criterion and have for testing each individual marker a statistical significance level of  $1/(\text{no. markers})$  times what it would be if only one hypothesis were tested (so called Bonferroni correction). When there are several (linked) markers available on the same chromosome, one should consider utilizing all the markers jointly and screen the whole chromosome interval by interval for existing of QTL.

Assume that the QTL is at the marker locus or in its immediate neighborhood. Therefore, it is possible to ignore how the recombination may reduce the estimate on the effect. When several linked markers are used, the localization part could be sharpened by the adjacent markers and subsequently, the estimation of QTL effect would receive better attention. Such an approach is called interval mapping. It was first developed for analyzing inbred lines [26].

Suppose there are now two markers  $M$  and  $N$  to test a case where  $B$  locus affecting the quantitative trait

variation is assumed to be in the interval between them. Further, the alleles at  $B$  recombine with the alleles of  $M$  at the rate  $c_1$  and with  $N$  with  $c_2$ . If the parental lines  $P_1$  and  $P_2$  are homozygous  $M_1 B N_1/M_1 B N_1$  and  $M_2 b N_2/M_2 b N_2$ , respectively, then in the backcross to  $P_2$  the expected progeny mean computed with respect to inheritance probabilities at the  $B$  locus is

$$\begin{aligned} & \text{prob}(B) \times \text{effect of } B \text{ allele} + \text{prob}(b) \\ & \times \text{effect of } b \text{ allele} = \text{effect of } b \text{ allele} + \text{prob}(B) \\ & \times (\text{effect of } B \text{ allele} - \text{effect of } b \text{ allele}) \end{aligned}$$

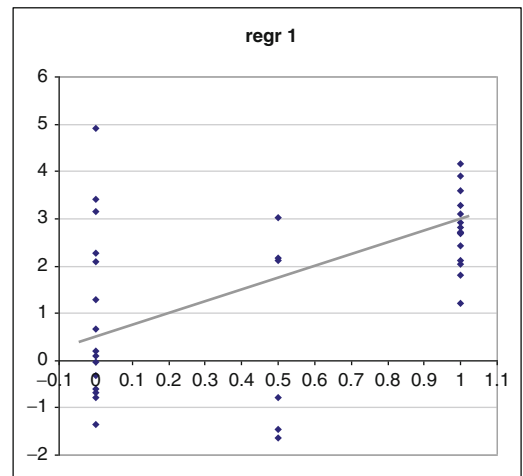
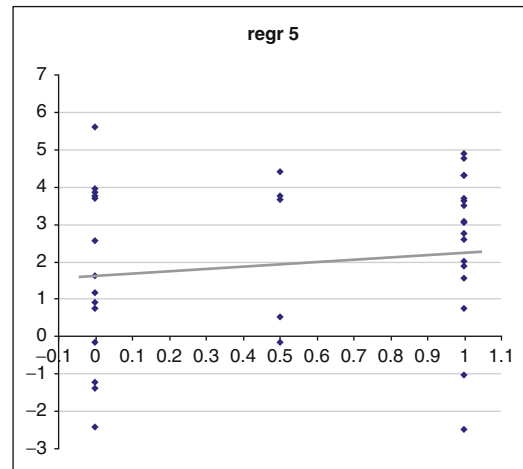
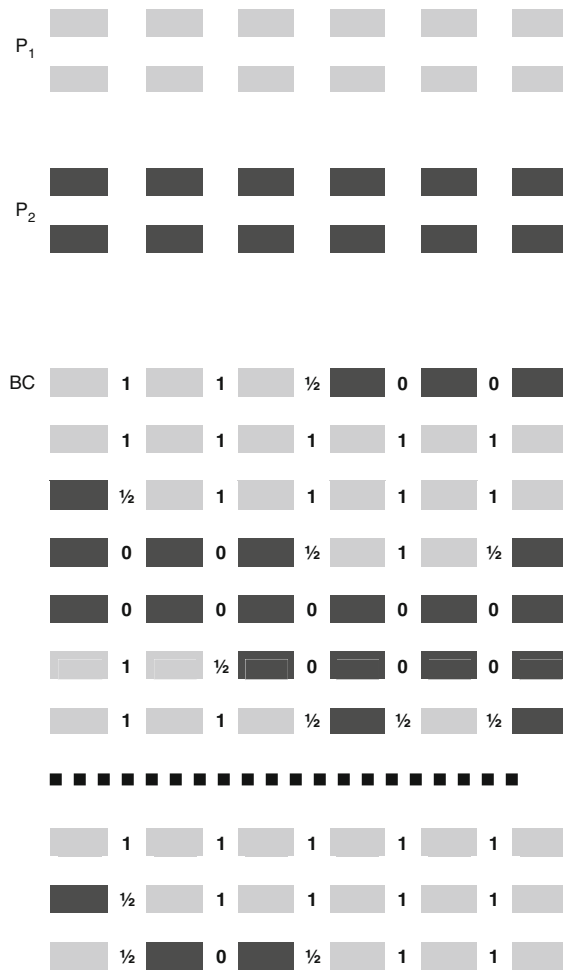
In terms of different combinations of  $M$  and  $N$  alleles in the backcross progeny, the probability for them having  $B$  allele from  $F_1$  parent is (recalling that the recombination rate  $c$  between  $M$  and  $N$  can be expressed as  $c_1 + c_2 - 2 c_1 c_2$ )

$$\begin{aligned} M_1N_1 & 1 - c_1c_2/(1 - c) \\ M_1N_2 & (1 - c_1)c_2/c \\ M_2N_1 & 1 - (1 - c_1)c_2/c \\ M_2N_2 & c_1c_2/(1 - c) \end{aligned} \quad (2)$$

The prob ( $B$ ) varies across the different positions between  $M$  and  $N$ . Different map positions could be converted to recombination rates in finding the map positions and allele effects that best explain the observations. The estimation could be done with computing the regression [27] at several positions between  $M$  and  $N$  with the model: performance record = intercept + regr coeff  $\times$  prob ( $B$ ) + residual. The regression coefficient is an estimate of the average effect of allele substitution at (biallelic) QTL. This is illustrated in Fig. 3. Like for any regression analysis, fitting the regression could be assessed with  $F$  test at different positions of the interval. If there are more markers along the studied chromosome, the analysis could be repeated over all the possible intervals.

When there is a reasonable coverage of the genome with interval mapping, the analysis is screening the whole genome for QTL. The same type of QTL analysis could be carried out for  $F_2$  cross animals. In that case also the dominance deviations could be estimated for QTL.

In farm animals, the closest to an inbred line is a breed. The first major genome scan utilizing breed



**Animal Molecular Genetics from Major Genes to Genomics. Figure 3**

QTL interval mapping via regression for backcross setting. The backcross (BC) population is produced by crossing phenotypically diverse "light" and "dark" parental lines ( $P_1$  and  $P_2$ ). The light line animals of  $P_1$  have a higher phenotype. The BC population is genotyped with a marker panel covering the shown genomic region at regular intervals of 20 cM. Markers are represented by shaded bars. The BC population is formed by crossing  $F_1$  animals with the  $P_2$  animals. When all the transmitted  $P_2$  markers are black, the BC animals are illustrated with the chromosomes from  $F_1$  parents, with possible recombinants. Marker intervals are considered one by one for the presence of QTL. The probability that individual has received the "light" QTL allele from  $F_1$  parent is computed (values in the brackets between the markers) and used as an explanatory variable to compute the regression of phenotypic value on it. The regression coefficient provides an estimate of the allele substitution effect. The QTL is in the first bracket and leads to higher regression coefficient than in the last (5th) bracket, far away from the QTL, both examples shown on the right. The illustration is modified from Georges [28]

differences was reported in the mid-1990s in pigs from a cross between the Large White breed and the wild boar [29]. Large QTL effects were found influencing both growth and fatness. After that study, several genome scans have been developed involving the commercial and exotic breeds.

In random bred populations, there is less regularity in many respects and the interval analysis should be preceded by the examination of allele phases across adjacent markers in heterozygous (say  $M_1/M_2N_1/N_2$ ) sires. The phase determination would require a large genotyped half-sib family from the sire. The non-informative

heterozygous progeny  $M_1/M_2$  cannot be automatically used for the analysis but the considered interval should be extended until an informative locus in the progeny is found. If the sire chromosome phases over the markers are  $M_1N_1/M_2N_2$  and there is a progeny that has inherited the combination  $M_1N_1$  (and assuming that the QTL allele  $B$  is on the same chromosome with  $M_1$  and  $N_1$ ), the progeny has  $B$  allele with probability  $1 - c_1 c_2/(1 - c)$ . The other probabilities could be derived and would resemble the ones listed for the backcross case above at (2).

The interval analysis is again performed within a family. The power of the analysis could be improved (or residual variation reduced) by analyzing several families and pooling the results at the end. The statistical model for progeny mean is therefore average level + sire<sub>*i*</sub> + regr coeff<sub>*i*</sub> × prob ( $B$ ) + residual. The sum of squares due to the regression is then obtained by pooling the sum of squares due to regression across sires as  $\sum SS_{\text{regr}}/\text{number of sires}$ . The degrees of freedom in the  $F$  test for the residual variation are the total number of observations minus twice the number of sires. The analysis is repeated at several positions, say at cM intervals, between  $M$  and  $N$  and further across other intervals. The highest value for the test parameter is indicating the likely position for QTL with the regression coefficient itself giving the substitution effect. The first genome-wide screening for QTL was done in the outbred half-sib design for dairy cattle by Georges et al. [30].

### Design of QTL Mapping Studies

It is important to understand what would be an optimal design and number of observations ( $n$ ) for QTL mapping studies. First consider the regression with a single marker for a crossbred population (or single family). The explanatory variable ( $prob$ ) is made of probabilities for the inherited QTL allele. For simplicity, there are so many observations ( $>50$ ) that one can formulate the test for the regression coefficient as regr. coeff./its standard error =  $\text{regr coeff}/(\sigma_e^2/SS_{\text{prob}})^{1/2} = \sqrt{n} \sigma_{\text{prob}} \text{ regr. coeff.}/\sigma_e$ . In other words, the test parameter depends on the number of observations, variation range of explanatory variable, size of QTL effect and residual variation ( $\sigma_e$ ).

Considering the type I error rate  $\alpha$  and test power  $1 - \beta$ , then in terms of the respective standardized

variates ( $z$ ) of normal distribution, the number of observations required to detect a QTL with an effect  $a/\sigma$  can be deduced from  $(z_{1-\alpha} + z_{1-\beta})^2/(\sigma_{\text{prob}} a/\sigma)^2$ . For the 5% risk and power 90% with the effect sizes

|     |     |     |     |
|-----|-----|-----|-----|
| 0.1 | 0.4 | 0.7 | 1.0 |
|-----|-----|-----|-----|

the required number of animals (with observation and marker genotype) in a backcross is

|       |     |    |    |
|-------|-----|----|----|
| 4,202 | 262 | 85 | 42 |
|-------|-----|----|----|

Compared to backcross, the explanatory variable would double its variation in  $F_2$  and therefore the  $F_2$  cross would have twice the efficiency given the same number of genotyped individuals. So far it is assumed that in the single marker analysis, the marker is at QTL itself. When this is not so and the recombination between marker and QTL is  $c$ , then with the same criteria the number of genotyped individuals should be  $1/(1 - 2c)^2$  times higher (for more details see [31]).

In conclusion, a cross between divergent lines is very powerful to detect QTL because the statistical power is high and a reasonable number (200–300) of animals is sufficient to detect a major gene. Experimental cross is also a useful approach for QTL screening as the production could be controlled in a small scale study and uncommon traits (disease resistance and quality traits) can be monitored. However, in farm animals, uniform lines are not common and rarely inbred enough to fully exploit the crossbreeding advantages. It is also very costly to maintain such lines or find facilities for generating and maintaining experimental crosses.

The random bred populations should be analyzed pooling within family results across families. When there are several groups in the analysis, the testing should follow the framework for the analysis of variance ( $F$  ratio). The test is improved with a large number of progeny. When there are  $s$  sires, the family size requirement could be computed from  $F^2 \sigma_e^2 s / \sum SS_{\text{prob}}$ . On the other hand, the more sires there are, the higher are the chances of finding a family with a segregating QTL.

With the half-sib design, the number of genotyped animals could be reduced by moving from daughter design to granddaughter design [32]. Instead of



genotyping daughters for markers, the progeny tested sons of the sires are now typed and their evaluation results are used in the analysis. Thereby the number of genotypings can be kept reasonable and the accurate genetic value of sons reduces the residual variation and thereby increases the power of analysis. The sons are genotyped for markers that are heterozygous in their sire. The analysis for a single marker is carried out within sires with the model: progeny test result value of son  $j$  from sire  $i = \mu + \text{sire}_i + \text{son's genotype}_{ij} + \text{residual}$ .

With a given number of genotypings the test would be more powerful than the daughter design. When there are 200 daughters per sire versus 200 sons per sire and 100 daughters per son, one can compare the test power (risk 5%) with different effect sizes for daughter/granddaughter design.

| Number of sires/paternal grandsires | No. genotypings | $a/\sigma = .1$ | $a/\sigma = .3$ |
|-------------------------------------|-----------------|-----------------|-----------------|
| 5                                   | 1,000           | 3/48            | 50/97           |
| 10                                  | 2,000           | 5/73            | 76/99           |
| 20                                  | 4,000           | 7/95            | 95/99           |

In order to reach the power of granddaughter design, the daughter design would need large progeny groups which – in terms of collecting DNA samples and carrying out genotypings – are workwise almost impossible.

For the interval mapping, both the variance of the explanatory variable and residual would be reduced while in general the aim should be to push down the residual variation as low as possible while keeping a wide range of alternatives for the explanatory variable. If only animals with extreme phenotypes are genotyped, the latter could be increased and the efficiency enhanced. Ultimately, genotyping only the very highest and lowest phenotypes, the variance of the explanatory variable is increased, and in the best situation the analysis would have only the extreme homozygotes for QTL. It is important to notice that the selective genotyping should be done within families to make sure that the selection is targeted on the QTL instead of the general genetic background. Selective genotyping has its drawbacks, as it would focus on a single trait at a time.

## Genes Underlying QTL

In farm animals, the QTL mapping has been very popular since the microsatellites and later SNPs became available as mapping markers. Depending on the species, two designs have been used: crosses mainly in pigs and chicken and half-sib families in dairy cattle. The QTL investigations have been carried out with marker sets of 100–300 microsatellites covering the whole genome. This would correspond to a marker density of some 10–30 cM. QTL mapping in farm animals has been very prolific as it has been possible to have sufficiently large experiments or population samples to yield an adequate statistical power for QTL detection. Until today, thousands of QTL for numerous traits have been reported (see <http://www.animalgenome.org/QTLdb/>).

| Year | Number of QTL (number of traits in brackets) |             |             |
|------|----------------------------------------------|-------------|-------------|
|      | Cattle                                       | Pigs        | Chicken     |
| 2006 | 630 (89)                                     | 1,287 (246) | 657 (112)   |
| 2011 | 4,682 (376)                                  | 6,344 (593) | 2,451 (248) |

## Fine Mapping

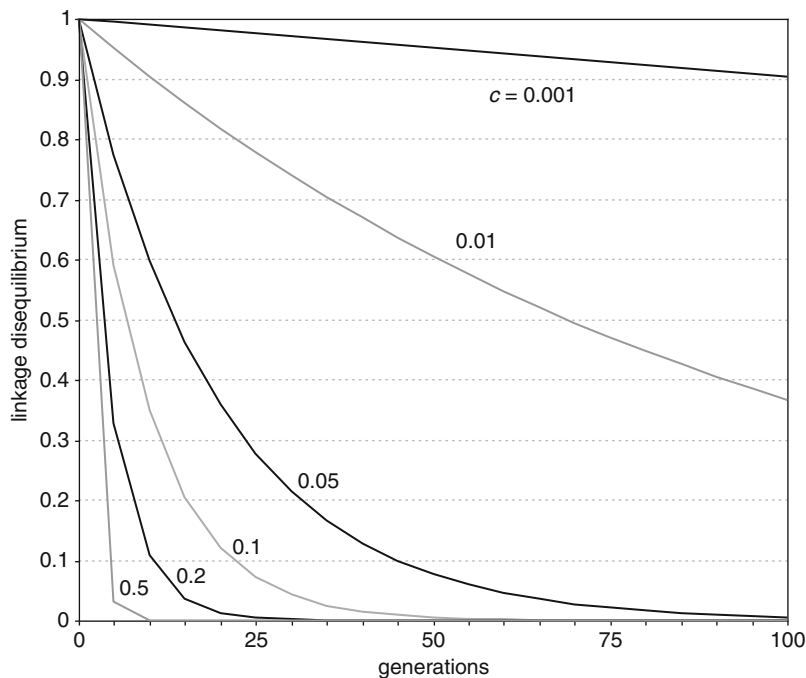
A successful *marker-assisted selection* requires that the applied marker is at the QTL itself or very close to it. Otherwise there is a need to resort to linked markers which may differ across families, depending on how they are showing *linkage disequilibrium* with the targeted QTL. There are other reasons for the interest to find the underlying genes, as the general aim is to understand more about their function and utilize them in further studies and applications. It is hard to resist the temptation of trying to find the genes, the allelic variants within them, the allelic effects and frequencies, co-effects with the alleles of other loci, and effects on other traits. Moreover, the never-ending curiosity drives research to distinguish coding areas and regulatory elements (e.g., [33]). In a typical QTL study, it is possible to locate a QTL only within a fairly wide region of a chromosome. A region of 10–30 cM may still contain 100–300 genes which is making difficult to identify and characterize the responsible mutations [28].

There are several factors that affect fine mapping. An obvious criterion is high marker density. The design consideration showed how the sample size of animals contributes substantially to the mapping resolution. As high as 5,000 individuals are required to increase the resolution down to 5 cM in crossbreeding designs. Therefore, it is necessary to think about ways to produce new recombinants. It is easier with experimental laboratory strains serving as model organisms where with a short generation turnover it is possible to have intercross populations  $F_3$ ,  $F_4$ ,  $F_5$ , . . . In that context, by generation  $t$  the confidence interval for the QTL location is reduced to  $1/t$  of the  $F_2$  one given the same number of animals per generation [31]. In farm animals, such operations would be money and time consuming and practically unfeasible.

It is easy to produce a large number of markers for an interesting QTL area and find the markers with the best association with QTL alleles. Usually, combining individual marker variation over the area and assessing the association between haplotypes and QTL alleles may be more fruitful than analyzing single markers

one by one [34]. In farm animals, the density of observable recombinants could be increased by exploiting historical recombination events in comparing the same genomic region within population or even across breeds (Fig. 4). Many breeds are originating from a small number of founders or have experienced bottlenecks alternating with periods of population growth. These events affect sampling of haplotypes and thereby cause *linkage disequilibrium* which can be used in fine mapping QTL regions. *Linkage disequilibrium* extends over longer genome stretches in young populations which hampers localizing causal mutations. In modern dairy breeds, *linkage disequilibrium* is seen as long haplotype blocks [35]. The older haplotypes shared by breeds can be exploited in high-resolution mapping.

In a QTL analysis one is following the transmittance of the variants over a chromosome section and computing probabilities for assumed QTL in the marker bracket. The more detailed variation within the chromosome interval is ignored and further it is assumed that the variants of the interval are unrelated. The intervals could be filled with more markers, especially



**Animal Molecular Genetics from Major Genes to Genomics. Figure 4**

The decay of linkage disequilibrium over generations for different values of recombination rate ( $c$ ) between loci. It is noticeable that with a recombination rate 1% it takes some 80 generations to halve very high linkage disequilibrium

around the found QTL, to classify the relationships between the interval variants given the linkage disequilibrium. When there is evidence for a QTL from linkage analysis and from *linkage disequilibrium* around the putative QTL position, the linkage and *linkage disequilibrium* analysis could be combined [36]. The analysis would eliminate QTL effects from distant regions of the chromosome and other chromosomes because they do not appear as linked QTL. Therefore, the combined analysis would produce a clearer signal for the QTL position compared to the two separate analyses [37]. Such a combined linkage and *linkage disequilibrium* approach has been successfully exploited in cattle to map QTL within less than 1 cM [38].

The discussion above shows how important it is to eliminate the spurious associations due to linked genes or other genetic background. One approach is to use transmission disequilibrium test or TDT [39]. This compares the effect of the two alternative types of gametes from a common parent. For example, a sire heterozygous for a studied marker (or a putative gene) may have a large progeny group allowing a test to find how regular the difference is between progeny inheriting the alternative alleles. Such an analysis is extended over sires by treating them fixed.

### Few Examples on Causal Mutations

The characterization of mutations underlying QTL is very hard in any species, and it is not too surprising that there are only few validated examples on such mutations in domestic animals. The work has usually started from a genome-wide screening for QTL and then proceeded to find genetic and functional support for possible causative mutations.

Many independent QTL mapping studies in dairy cattle have shown that there is QTL for milk traits on chromosome 14. After lots of effort and international collaboration it has been shown that the associations could be explained by the *DGAT1* gene that is coding acyl-coenzyme A: diacylglycerol acyltransferase. The enzyme is known to catalyze the last step in the triglyceride synthesis and a missense mutation (K232A) influences milk fat content in cattle (e.g., [40]). The fine mapping and gene identification took some 7 years.

The second example is from pigs. A single-nucleotide substitution of *IGF2* (insulin-like growth

factor 2) is a causal gene behind a major QTL for growth [41]. This is a regulatory mutation. It has been present in four different breeds selected for lean growth and a major reason for the successful detection is the finding of an ancestral haplotype that differs only by one nucleotide substitution from the mutant haplotype [42].

Another example is a single nucleotide substitution in the myostatin gene (*MSTN*) causing increased muscle mass in Texel sheep [43]. This mutation creates a new target site for two microRNAs expressed in muscle which leads to down regulation of *MSTN* transcripts.

The findings have a clear message: farm animal populations have a high haplotype diversity allowing successful genetics research to locate genes underlying QTL irrespective of the mode of action. The examples show that the molecular nature of a QTL can be as simple as a missense mutation in a coding sequence or a single nucleotide substitution in noncoding DNA. It is also impressive how in farm animals the genes behind the found QTL are regulatory mutations. The regulatory factors are much more challenging in showing the effect on gene function than mutation changing the protein sequence and structure.

There are others reported but in many cases more work is needed to demonstrate their effect in independent samples. The number of actual genes in the QTL seems very low so far but is comparable with the modest findings in man and mice where the work volume is more substantial. One obvious reason for a small number is the poor resolution in QTL mapping experiments. The confidence interval for a QTL is at best very wide and may contain hundreds of genes. Another reason is that many QTL will not have such a simple explanation as the examples described above. Some QTL will be due to several mutations present in one or many genes. Such QTL may break up into several linked loci for the fine mapping. Mutations in noncoding DNA make their detection and functional characterization also very difficult (see [28, 42] for more discussion).

### Candidate Gene Approach

In the characterization of genetic factors, there are basically two types of approaches. The linkage studies

are aiming to locate genomic regions or QTL that are important sources of variation, or biological knowledge is used to deduce the genes most likely responsible for the variation. The examples above are indicating that studies to identify positional *candidate genes* have a rather limited success. It is not therefore surprising that many researchers have restricted their studies to investigate the role of biologically deduced *candidate genes* in explaining the variation in the available data. This approach is much more economical, on the other hand there is very little new knowledge to be discovered as the analysis only spotlights the chosen gene or genes. Sometimes, the studies have switched to candidate approach after the linkage approach has failed or included *candidate genes* as a marginal part of wider analysis not originally designed to explore individual genes. The research angle is naturally highly subjective in choosing candidates from numerous promising ones.

The report on functional candidates should provide full details why the particular genes were chosen, where the genotyped group of animals is originating, what kind of marker set was used to eliminate the other associated effects in the genome, etc. For any gene, the combined data on its vicinity should be used to convince others about hitting the true gene in the face of too much interfering information from dozen other genes. A good design in such studies should be aiming at balanced genotype frequencies. If a gene is picked up because of ease of genotyping, the only hope is to have much luck to hold marker variants that are in *linkage disequilibrium* with the alleles causing a trait deviation. There is also the trick of having a “significant” result surface from the data by permuting several genes and traits (see [44, 45]).

It is a common problem that *candidate genes* are not providing consistent evidence when exposed to validation in subsequent generations of the same population or more importantly, in other populations. The most beautiful investigations are finalized by functional mutation studies (knock-out or in) or complementation tests in a modeled mouse experiment, practically impossible in livestock species. One way to improve the analysis is to define a trait more explicitly. For example, calving interval could be itemized to several sub-traits, which in turn may prove to be more rewarding objects for QTL analysis. Likewise with the markers, instead of

associating gene variants, it is now possible to go a level deeper and excavate the respective gene expression patterns. The difficulties are not over when it comes to deciding which tissues and which developmental stage would provide appropriate samples for expression studies and how to plan a comparable and uniform production environment for such studies. The expression profiles are obviously an outcome influenced by several factors and their interactions are ever so hard to interpret (for more discussion, see [28]).

The complex traits in medical studies have experienced similar disappointing conclusion, and there are hardly any major genes manifesting the familial multifactor diseases. Researchers have voiced this by talking about missing explanations for the observed genetic variation or heritability [46]. In quantitative genetics, it is more and more obvious that the variation is simply caused by lots of genes with tiny effects and variants that are not appearing among the prior candidates [47]. With the rapid growth of high throughput marker typing and re-sequencing, there is now a possibility with very large population samples and concerted actions to detect reliable associations even for the small effects [48]. Different techniques, such as detailed mapping and functional studies, could be combined [49]. What remains is the requirement for large bodies of data, as has always been in analyzing genetic variation for animal breeding purposes.

## Genomic Revolution

The very fast development of molecular genetic techniques has resulted in exploitation of the existing nucleotide diversity. The genome sequence is available for chicken [50] and cattle [51]. The sequences of the whole genome in pig and sheep are still on the way. The sequence information is indicating that a typical animal genome has almost endless amount of potential variable sites. Several descriptive measures are used to summarize polymorphisms of DNA sequences. Under a neutral model, the expected level of diversity can be deduced from the generation of new alleles by mutations and from the elimination of alleles by drift (which is inversely proportional to effective population size), that is,  $4 N_e \times \text{mutation rate}$ . For comparison, in a human population two randomly chosen individuals differ at  $\sim 1$  in 1,000 nucleotides (1 SNP per kilo base).

The genetic diversity of mankind is low compared to other (older) species. In cattle and sheep the mean nucleotide diversity is 2–2.5 SNPs per kilo base [52], whilst for chicken the estimate is 4–5.5 [53]. With modern DNA chip technology up to hundreds of thousands of loci across the genome could be used to genotype animals with a reasonable cost. These possibilities have opened up completely new kind of possibilities to understand and utilize the genetic variation.

### MAS with More Markers

When there are several markers available, the *mixed model equations* for *breeding value* estimation could be extended to accommodate several markers with a vector of QTL effects. Using the previously introduced notation (Fig. 5), the use of several markers is with matrix presentation  $y = X\beta + Za + Qq + e$ , where  $Q$  is the incidence matrix for  $q$  [54].  $\text{Var}(q) = G\sigma_q^2$  where  $\sigma_q^2$  is the variance due to QTL and  $G$  is the matrix of probabilities for QTL alleles being identical by descent. These probabilities can be computed from the pedigree, marker, and linkage map information. So far QTL mapping has produced fewer useful genes or markers for MAS than was anticipated when the QTL screening work started. The found QTL are contributing much less to the variation than would be satisfactory for a successful use of MAS, as its efficiency is affected by the proportion of variation due to a major locus [55]. Much caution should be also taken in MAS, as most of the found QTL have an overestimated effect (so-called Beavis effect). The availability of vast amounts of SNP markers should improve both these aspects: a dense marker panel covering the whole genome would be able to wrap all the genetic variation and avoid the problems of few overemphasized markers.

### From MAS to Genomic Genetic Values

For the future practical application in estimating the genetic values, the most promising approach seems to be simultaneous utilization of a vast number of markers over the entire genome [56]. The map density is so high that the recombination between markers and QTL can be ignored. With a genome-wide set of markers, the estimate of individual's genetic value or genomic estimated *breeding value* is obtained by

*conventional selection*  
additive genetic values

$$y = X\beta + Zu + e$$



*marker assisted selection*

$$y = X\beta + Zu + Qq + e$$



*genomic selection*

$$y = X\beta + Mg + e$$

$u$  (polygenic) additive effects  
 $q$  QTL effects  
 $g$  marker effects  
 $e$  residual

$\beta$  fixed non-genetic factors

incidence matrices for

$X$  fixed effects  
 $Z$  additive effects  
 $Q$  QTL effects  
 $M$  marker effects

### Animal Molecular Genetics from Major Genes to Genomics. Figure 5

The mixed model methodology is a flexible tool to accommodate the models ranging from simple breeding value prediction with polygenic model to QTL effects associated with known markers. The effects of genome-wide marker sets are computed through a straightforward summation over the genome.  $X$ ,  $Z$ ,  $Q$ , and  $M$  are the incidence matrices linking the observations to respective effects

estimating effects associated with the markers from animals with both the phenotype and marker genotype information and summing the marker effects for recently genotyped animals available for selection. With genome-wide marker panel one can hopefully catch most of the genetic variation and produce a more attractive way to exploit markers.

The process of predicting the *breeding values* utilizing genomic information could be described in three steps:

1. Use a genome-wide marker panel to genotype animals. It is now a norm to use densely mapped SNPs as markers. The SNP variation is seen as two alleles (potentially with four). When the marker panel is very dense and genome-wide and each marker is wrapping the genetic variation for a measured trait in its neighborhood, the whole panel is covering all the genetic variation. With additive gene action the biallelic marker is assumed to have a substitution



effect. As the whole genome is covered with the dense map, the effects of the  $n$  SNP loci would sum the genetic value of the animal.

2. Estimate the effects of each marker locus. The model for the phenotype of animal  $i$  is  $y_i = \sum_{j=1}^n m_{ij}g_j + e_i$  where  $m_{ij}$  (2, 1 and 0) is representing the marker genotypes in the  $i$ th animal,  $g_j$  is half of the substitution effect at the marker locus for the quantitative trait and  $e$  is the residual. The phenotype and genotype data could therefore be used to estimate the marker effects (ignoring all the nongenetic factors) with the model in matrix notation  $\mathbf{y} = \mathbf{M}\mathbf{g} + \mathbf{e}$ . The new method could be seen as part of the development in mixed model methodology (Fig. 5). The first two steps are carried out in a sufficiently large reference population with information on both marker genotypes and trait phenotypes.
3. The genomic *breeding value* could be computed for genotyped selection candidates by summing the estimated effects over the markers.

There are several options how the equations could be solved. Treating the marker effects as random, it is customary to assume that their variance is constant over the genome or that there are few loci with large effects with the majority having small effects. In addition, there are options for the distribution of the effects. The case of normally distributed effects with constant variance is interesting because the *breeding value* predictions would be equivalent with *BLUP* where the pedigree-based relationship matrix  $A$  is replaced by a relationship matrix estimated from the marker information [57]. The latter is also called genomic relationship matrix and is technically *MM*. The genomic information has also been successfully used to incorporate the realized relationships into the  $A$  matrix used in *BLUP* [58].

It is the Bayesian methods that are able to cope with situations where an allowance for graduated effects of markers is made. This is beneficial as the QTL screening works is clearly showing that there are genomic regions having relatively large effects. Many distributions have been tried: a case where most markers have a small effect and very few a large effect, alternatively assuming 0 for several markers and nonzero for few markers.

These were studied in the original paper [56] and the former approach is known as BayesA and the latter one as BayesB. Meuwissen and Goddard [36] also presented an alternative where the variances of marker effects are sampled from a mixture of two distributions allowing the variation in the effect size across markers. Simple methods are possibly favored so far, as pure Bayesian approaches would require more computing time. Related to this, the rapidly increasing number of markers – hundreds of thousands – and availability of sequence data and larger number of genotyped animals would also challenge the computing capacities.

The accuracy of genomic *breeding values* depends of course on the number of records and proportion of genetic variation or heritability  $h^2$ . Small *effective population size* is generating linkage disequilibrium. Related to linkage disequilibrium, the genome contains regions of reduced haplotype diversity, termed haplotype blocks [59], separated by blocks of higher diversity. Within such blocks the frequencies of marker alleles may be highly correlated across loci. The genome blocking sets limits to the estimation even if the marker density is very high. The number of independent segments for a chromosome of length  $L$  (in *Morgans*) is  $2 N_e L / \ln(4N_e L)$  [57]. So, a chromosome of 1M has some 33 independent segments and the whole genome of 30 such chromosomes would have 1,000 independent segments. If the *effective population size* is small, increasing more markers would not improve the accuracy or resolution in explaining the genetic variation and the best one can do in that situation, is increasing the number of animals in the analysis [57, 60]. In conclusion, the accuracy of genomic *breeding values* would depend on the number of animals in the reference population, the heritability of the trait, the marker density, the number of independent genome segments (or effective population size) and the total length of the genome.

With best linear unbiased prediction and Bayesian method, the accuracy of predicting genetic values for offspring of the recorded animals has been shown to be of the order 0.7–0.8 [56], comparable to that of the progeny test. The genomic selection method was presented in 2001 and did not receive much attention, as at that time the required coverage of the whole genome with markers was not feasible, at least in terms of costs. When the high throughput genotyping

of SNPs with microarrays became available and Schaeffer [61] pointed out the remarkable returns and savings that the genomic selection could produce in dairy cattle breeding, the breeding companies rushed out to exploit the tool with unparalleled enthusiasm. He compared genomic selection with conventional progeny testing and assumed that there were marker effects available with accuracy 0.75 to select bulls and bull dams. As much less bulls would be kept after genomic “pre-selection,” the cost of the marker-based improvement would be only 10% of the traditional one. Another radical feature of genomic selection, is the sufficiently reliable evaluation of young animals which would reduce the generation interval drastically, for example, in the bull–bull path from 6 years down to less than 2 years. When it takes 5–7 years to have the progeny test result for a dairy bull, the genomic genetic value is available at birth. Putting the slightly lower accuracy and shortened generation interval together, the annual genetic change is predicted to be doubled.

Much of the work on assessing the advantages of genomic selection has been done by simulation. The reliabilities of genomic *breeding values* could be assessed by dividing the real data into two parts, usually over birth date and compare how the predictions based on older animals compare with the estimates obtained for the more recent ones with all the data. VanRaden et al. [62] concluded that the reliability is proportional to the size of the reference population while the increments in marker density had a smaller impact. They were implying that a reference population of few thousand progeny tested bulls is required to arrive at a satisfactory accuracy of genomic *breeding values*. Only the Holstein breed has such numbers and even there the breeding companies would need collaborative efforts across countries for reaching a sufficient number of accurately known sires.

Breeds with smaller populations (and progeny tested bulls) could in principle resort to the marker effects estimated in a large population. If the marker density is very high, genomic *breeding values* estimated in one breed may support the estimation in the other breed [63]. They may be some common *linkage disequilibrium* within a short distance across the breeds otherwise recombination has broken associations since the divergence. Selection and drift have also changed the genotype frequencies over time.

## Management of Variation

Very intensively selected populations may suffer from losing variation or, in a more extreme case, from inbreeding depression and appearance of recessives defects. This is due to the reduced number of selected parents to generate the breeding bulls. The risk for inbreeding and drift is expressed with *effective population size* or rate of inbreeding ( $\Delta F = 1/(2N_e)$ ). The safe areas for these would be over 50 and less than 1%, respectively (e.g., [64]). In dairy cattle, the awareness on rate of inbreeding was triggered by the breeding schemes utilizing multiple ovulation and embryo transfer (MOET). It was shown that the schemes could be optimized with respect to genetic progress and costs, by concentrating the MOET operation on cows in a nucleus of few hundred cows only. The apparent small number of parents and larger full-sib families accompanied by early selection would easily mean risks for higher rates of inbreeding. This kind of worries generated lots of research and now the better understanding about the related risks in selection programs has resulted in new tools. The most important aspect is to maximize the variation in selected groups of males and females while the formation of mating pairs is less relevant. Meuwissen [65] has developed so-called optimum contribution method which yields the guidance in choosing parents and their progeny numbers to maximize the genetic gain while minimizing the coancestry. The constraining on the coancestry is based on the relationship matrix and therefore the optimum contribution method could be easily integrated with the information needed for *BLUP* evaluation.

Genomic *breeding values* are based on information on the individuals themselves and are therefore better than *BLUP breeding values* which may more often lead to co-selection of sibs due to family information [66]. So the genomic selection has short-term advantages for the management of genetic variation. In the long-term an efficient genomic selection is automatically reducing variation. The genomic information is on the other hand giving an accurate estimate on the relationship between individuals. While the pedigree based matrix gives the expected relationships, the marker-based matrix yields the realized relationships. The earlier work (e.g., [67]) showed that tens or even hundreds

of markers (e.g., microsatellites), is of limited value in estimating the relationships and in controlling the rate of inbreeding. The use of genome-wide SNP panels has changed the situation and molecular similarities could be calculated very accurately and used like the genealogical information. Just like the relationship matrix in the *BLUP* context is improved by the genomic information, in the same way the genomic relationship matrix could enhance the management of variation. There are only preliminary results [68] which are indicating that high benefits could be anticipated from the use of genomic relationship matrix in promoting the sustainable use of genetic variation.

### Genetic Correlation

It is a common observation that there is a negative genetic correlation between production traits and longevity, made of such traits as disease resistance and fertility. It is widely accepted that such a correlation is caused by pleiotropy so that the same genes would have opposite effects on the two sets of traits. This could be understood as a consequence of allocating limited metabolic resources to competing physiological compartments within an animal. If a genetic correlation between two traits is positive, it can be thought that the effects on the two traits are on average influencing to the same direction. On the other hand, if the majority of effects are of opposite “sign,” a negative correlation would be seen. However, there should be some variation over the pleiotropic genes in the genome, there being also some with positive effect on both the traits. Therefore, selection could be differentiated over the genome to avoid undesirable effects in longevity traits while improving the production efficiency. This is something one is going to learn when multiple trait analysis are thoroughly investigated with the dense genome-wide marker information.

### Extensions to a Pair of Populations

The introgression of multifactor traits from one population to another should enjoy the powerful genomic tools in making the procedure more precise and diminishing the compromises of receiving undesirable sections of the donor genome. Ødegård et al. [69] showed that genomic selection results in an efficient

introgression of desirable QTL alleles from a donor line, as the genome of the donor apart from the QTL could be selected against. Lots of data is obviously needed for separating the beneficial and unattractive genome parts in a candidate exotic breed.

*Heterosis* or hybrid vigor in crosses has been long observed. Predicting which lines will give good crosses is a hit and miss affair. Simple dominance is sufficient to give crossbreds better performance than either parent provided that both parents are fixed for the dominant allele at some locus at which the other parent is fixed for the recessive one – that is with one parent being  $Bc/Bc$  and the other  $bC/bC$ , where  $B$  and  $C$  are the dominant alleles.

Several breeding procedures have been suggested and used in order to make best use of the *heterosis*, among them the reciprocal recurrent selection, proposed by Comstock et al. [70] makes the most use of quantitative genetic principles. The benefits from crossbreeding are highest with widely deviating allele frequencies between the breeds. There is a need for efficient methods with immediate returns to evaluate the most promising breed crosses jointly with the most potential mating pairs.

Starting from the results by Smith and Mäki-Tanila [6], theory and methods to compute genotypic means and covariances in a two-breed population under dominance inheritance have been presented by Lo et al. [71]. They showed that the genotypic mean is a linear function of 5 location parameters and that the genotypic covariance between relatives is a linear function of 25 dispersion parameters. These would include the additive (and the corresponding heritability) and dominance variance in both the purebred populations and the variation of contributions from the breeds to the  $F_1$  individuals. Clearly simpler and more parsimonious methods are needed.

Genomic research has proven to be a powerful approach in quantifying the genetic distances between populations, in revealing history of animal populations, number and sites for domestication, population expansions and contractions, selection, origin, and mixing of maternal and paternal lineages (see [72]). Because *heterosis* is proportional to the differences in gene frequencies in the parental lines, it is possible to make marker-based prediction of hybrid performance based on genetic distances, despite having

only indirect estimates of allele frequencies for the interesting traits *via* the anonymous markers.

Xu [73] stimulated by the elegance of genomic selection has extended a Bayesian analysis to the  $F_2$  population of inbred lines and has successfully also estimated the dominance deviations. However, the cases starting from random breeding populations – in subdivided population or in crossbred animals, typical in animal breeding – need research. Toro and Varona [74] have investigated the use of genomic information in predicting also dominance effects and the advantages of designing matings to have a full advantage of the extra component in the evaluation. First of all, the genomic approach was possible compared to the overcomplicated polygenic parameterization. Secondly, the prediction of additive effects had a higher accuracy. Finally, the immediate response was higher when dominance effects were estimated, although the subsequent progress stayed very much the same. If the cost of genotyping is reasonable, the estimation of dominance effects could be seen as a new tool in utilizing *heterosis* in chicken and pig production where crosses of different lines are a norm.

### Research on Beneficial and Harmful Genes

Along the use of genome-wide marker sets in large reference populations, it is possible to map medium to large size QTL. This would yield genomic predictions that are stable across families and generations. If the genes underlying QTL are found, they may be further studied for the pleiotropic effects on other traits, dominance effects, interactions with other QTL (epistasis), or response to environmental changes. Information derived from such studies will lead to better models and better predictions and even management of phenotypes, which can be used for selection and production planning.

There are several Mendelian defects identified in farm animals. Recessive disorders are a problem in efficient improvement schemes, because healthy carriers can spread the disease allele quickly to a large number of progeny. A good example is BLAD (bovine leukocyte adhesion deficiency) causing a severe immunodeficiency. The mutation got widely distributed in the Holstein cattle due to a famous bull few decades ago. The missense mutation is now identified [75] and

the eradication program has been successful. Charlier et al. [76] reported fine-scale mapping of five recessive disorders in cattle using large SNP panel. Between 25,000 and 50,000 SNPs were used in the discovery populations. Homozygosity mapping is used to detect mutations that cause disease when both copies are present (recessivity). Three disorders were mapped with a sufficient resolution so that the molecular basis could be characterized and effective eradication tools were established and used in mating planning. A prerequisite for finding recessive disorders is a well-organized recording of defects and diseases in the population. In general, molecular genetic techniques could be used to develop diagnostic tests. On the other hand the disorders are providing animal models to study human diseases.

### Future Directions

The conventional prediction methods based on *infinitesimal model* have produced very impressive results in quantitative traits of animal production. There is, however, plenty of evidence on the existence of major genes mediating the variation in such measured traits. For some time now, it has been possible to map such genes in the genome using molecular genetic markers. The accurate localization is needed for estimating the gene effects and integrating their direct selection into methods predicting the polygenic effects. There has been extensive genome screenings for QTL carried out all the farm animal species, either with crossbreeding or half-sib design. The outcome is thousands of QTL while the high-resolution localization has yielded only very few causal mutations underlying the found QTL. Also the variation contributed by these QTL is far too low for efficient marker- or gene-assisted selection. The molecular genetic technology has in the mean time taken further leaps and now there are available microarray chips containing thousands or even hundreds of thousands of SNPs to allow genomic selection. Genomic selection is in principle genome-wide *marker-assisted selection*. As it covers all the genetic variation in the genome, the estimated of marker effects would give sufficiently reliable predictions of genetic values for young animals. Hence, the genetic improvements programs could be accelerated and high savings could be done in testing schemes. Genomic information could be further used for management of genetic

variation, mating planning utilizing non-additive variation, understanding genetic correlation architecture, etc. The animal breeding industry has quickly adopted the new technology and is now widely exploiting the latest SNP panels in the selection schemes. As a side product, the systematic genome-wide screening is picking up harmful and beneficial mutations for further research and applications. It is very tempting and certainly very challenging to try to model the variation via complicated pathways and interactions due to individual genes and regulatory factors. No doubt many would try. Luckily animal breeding research has a long and flourishing history of sophisticated mathematics and statistics. Hence, much research will be devoted to the area. The main test for a new method in animal breeding would consist of questions like: how it is going to help us in understanding variation, how would it help to improve prediction of breeding value, what gains are made in efficiency of selection programs, what kind of savings could be made in testing, etc. Finally, to have firm ground in analyses and decision making, there is a continuous need to have lots of information, both on the genomes and the measurable traits.

## Acknowledgment

I am grateful to colleagues, especially Pekka Uimari, at MTT, for useful comments on the draft of the text, and to Ignacy Misztal for lots of encouragement.

## Bibliography

### Primary Literature

1. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, Harlow, 464 p
2. Weller JI (2009) Quantitative trait loci analysis in animals, 2nd edn. CABI, Wallingford
3. Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph
4. Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used for prediction of breeding values. *Biometrics* 32:69–79
5. Wray NR, Hill WG (1989) Asymptotic rates of response from index selection. *Anim Prod* 49:217–227
6. Smith SP, Mäki-Tanila A (1990) Genotypic covariance matrices and their inverses for models allowing dominance and inbreeding. *Genet Sel Evol* 122:65–91
7. Silvela L (1980) Genetic changes with generations of artificial selection. *Genetics* 95:769–782
8. Mäki-Tanila A (1982) The validity of the heritability concept. Ph.D. thesis, University of Edinburgh, 120 p
9. Mäki-Tanila A, Kennedy BW (1986) Mixed model methodology under genetic models with a small number of additive and non-additive loci. In: *Proceedings of 3rd world congress on genetics applied to livestock production* 12, Lincoln, pp 443–448
10. Webb AJ, Carden AE, Smith C, Imlah P (1982) Porcine stress syndrome in pig breeding. In: *Proceedings of 2nd world congress on genetics applied to livestock production* 5, pp 588–608
11. Merat P, Ricard FH (1974) Etude d'un Gene de nanisme lié au sexe chez la poule: importance de l'état d'engraissement et gain de poids chez l'adulte. *Ann Genet Sel Anim* 6:211–217
12. Rollins WC, Tanaka M, Nott CFG, Theissen RB (1972) On the mode of inheritance of double muscled conformation in bovines. *Hilgardia* 41:433–456
13. Piper LR, Bindon BM (1982) The genetic segregation for fecundity in Booroola Merino sheep. In: Barton RA, Smith WC (eds) *Proceedings of world congress on sheep and beef cattle breeding*. Dunmore, Palmerston North, pp 315–331
14. Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756
15. Dekkers JCM (2004) Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J Anim Sci* 82:E313–E328
16. Pong-Wong R, Woolliams JA (1998) Response to mass selection when an identified major gene is segregating. *Genet Sel Evol* 30:313–337
17. Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. *Nat Rev Genet* 3:22–32
18. Yancovich A, Levin I, Cahaner A, Hillel J (1996) Introgression of the avian naked neck gene assisted by DNA fingerprint. *Anim Genet* 27:149–155
19. Knap PW, Neeteson-Van Nieuwenhoven AMA (2005) Private and public roles in conservation. In: *Proceedings of options and strategies for conservation of FAnGR.*, Montpellier, pp 62–67
20. Bidanel JP (1989) Etude de stratégies de valorisation en croisement de la race Meishan. 3 – evaluation comparée de différents systèmes de croisement. *J Rech Porcine en France* 21:361–366
21. Legault C (1998) Génétique et prolificité chez la truie: la voie hyperprolifique et la voie sino-européenne. *INRA Prod Anim* 11:214–218
22. Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer, Sunderland
23. Elston RC, Stewart J (1973) The analysis of quantitative traits for simple genetic models from parental, F1 and backcross data. *Genetics* 73:695–711
24. Neimann-Sorensen A, Robertson A (1961) The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agric Scand* 11:163–196



25. Soller M, Genizi A (1978) The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics* 34:47–55
26. Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185–199
27. Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69:315–324
28. Georges M (2007) Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annu Rev Genomics Hum Genet* 8:131–162
29. Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M, Andersson K, Andersson-Eklund L, Edfors-Lilja I, Fredholm M, Hansson I et al (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* 263:1771–1774
30. Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR, Zhao X, Womack JE, Hoeschele I (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* 139:907–920
31. Darvasi A (1998) Experimental strategies for the genetic dissection of complex traits in animal models. *Nat Genet* 18:19–24
32. Weller J, Kashi Y, Soller M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J Dairy Sci* 73:2525–2537
33. Mackay TF (2009) The genetic architecture of complex behaviors: lessons from *Drosophila*. *Genetica* 136:295–302
34. Calus MP, Meuwissen TH, Windig JJ, Knol EF, Schrooten C, Vereijken AL, Veerkamp RF (2009) Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genet Sel Evol* 41:11
35. Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* 10:220–227
36. Meuwissen THE, Goddard ME (2004) Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet Sel Evol* 36:261–279
37. Goddard ME, Meuwissen TH (2005) The use of linkage disequilibrium to map quantitative trait loci. *Aust J Exp Agric* 45:837–845
38. Meuwissen THE, Karlsten A, Lien S, Olsaker I, Goddard ME (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161:373–379
39. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin dependent diabetes mellitus. *Am J Hum Genet* 52:506–516
40. Grisart B, Farnir F, Karim L, Cambisano N, Kim JJ, Kvasz A, Mni M, Simon P, Frère JM, Coppieters W, Georges M (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci USA* 101:2398–2403
41. Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, Andersson G, Georges M, Andersson L (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* 425:832–836
42. Andersson L (2009) Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. *Genetica* 136:341–349
43. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibé B, Bouix J, Caiment F, Elsen JM, Eychenne F, Larzul C, Laville E, Meish F, Milenkovic D, Tobin J, Charlier C, Georges M (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet* 38:813–818
44. Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 3:391–397
45. Zhu M, Zhao S (2007) Candidate gene identification approach: progress and challenges. *Int J Biol Sci* 3:420–427
46. Maher B (2008) Personal genomes: the case of the missing heritability. *Nature* 456:18–21
47. Hill WG (2010) Understanding and using quantitative genetic variation. *Phil Trans R Soc B* 365:73–85
48. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin N, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42:565–569
49. Karim L, Takeda H, Lin L, Druet T, Arias JA, Baurain D, Cambisano N, Davis SR, Farnir F, Grisart B, Harris BL, Keehan MD, Littlejohn MD, Spelman RJ, Georges M, Coppieters W (2011) Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat Genet* 43:405–413
50. Wallis JW, Aerts J, Groenen MAM, Crooijmans RPMA, Layman D, Graves TA, Scheer DE, Kremitzki C, Fedele MJ, Mudd NK, Cardenas M, Higginbotham J, Carter J, McGrane R, Gaige T, Mead K, Walker J, Albracht D, Davito J, Yang S-P, Leong S, Chinwalla A, Sekhon M, Wylie K, Dodgson J, Romanov MN, Cheng H, de Jong PJ, Osoegawa K, Nefedov M, Zhang H, McPherson JD, Krzywinski M, Schein J, Hillier L, Mardis ER, Wilson RK, Warren WC (2004) A physical map of the chicken genome. *Nature* 432:761–764
51. Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324:522–528
52. Meadows JRS, Hawken RJ, Kijas JW (2004) Nucleotide diversity on the ovine Y chromosome. *Anim Genet* 35:379–385
53. Hillier LW, Miller W, Birney E et al (2004) Sequence and comparative analysis of the chicken genome provide unique

- perspectives on vertebrate evolution. *Nature* 432:695–716 (International Chicken Genome Sequencing Consortium)
54. Fernando RL, Grossman M (1989) Marker assisted selection using best linear unbiased prediction. *Genet Sel Evol* 21: 467–477
  55. Meuwissen THE, Goddard ME (1996) The use of marker haplotypes in animal breeding schemes. *Genet Sel Evol* 28:161–176
  56. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
  57. Goddard ME (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136:245–257
  58. Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. *J Dairy Sci* 92:4656–4663
  59. Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 4:587–597
  60. Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3:e3395
  61. Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet* 123:218–223
  62. VanRaden PM, VanTassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* 92:2737–2746
  63. De Roos APW, Schrooten C, Mullaart E, van der Beek S, de Jong G, Voskamp W (2009) Genomic selection at CRV. *Interbull Bull* 39:47–50
  64. Woolliams J, Berg P, Mäki-Tanila A, Meuwissen T, Fimland E (2005) Sustainable management of animal genetic resources. *NGH Nordisk Genbank Husdyr*, Norway, 95 p
  65. Meuwissen THE (1997) Maximizing the response of selection with a predefined rate of inbreeding. *J Anim Sci* 75:934–940
  66. Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007) Inbreeding in genome-wide selection. *J Anim Breed Genet* 124:369–376
  67. Fernandez J, Villanueva B, Pong-Wong R, Toro MA (2005) Efficiency of the use of pedigree and molecular marker information in conservation programmes. *Genetics* 170:1313–1321
  68. Sonesson AK, Woolliams JA, Meuwissen THE (2010) Maximising genetic gain whilst controlling rates of genomic inbreeding using genomic optimum contribution selection. In: Ninth world congress on genetics applied to livestock production, Leipzig
  69. Ødegård J, Yazdi MH, Sonesson AK, Meuwissen TH (2009) Incorporating desirable genetic characteristics from an inferior into a superior population using genomic selection. *Genetics* 181:737–745
  70. Comstock RE, Robinson HF, Harvey PH (1949) A breeding procedure designed to make maximum use of both general and specific combining ability. *J Am Soc Agron* 41:360–367
  71. Lo LL, Fernando RL, Cantet RJC, Grossman M (1995) Theory for modelling means and covariances in a two-breed population with dominance inheritance. *Theor Appl Genet* 90:49–62
  72. Toro MA, Mäki-Tanila A (2007) Genomics reveals domestication history and facilitates breed development. In: Oldenbroek K (ed) *Utilisation and conservation of farm animal genetic resources*. Wageningen Academic, The Netherlands, pp 75–102, 232 p
  73. Xu S (2003) Estimating polygenic effects using markers of the entire genome. *Genetics* 163:789–801
  74. Toro MA, Varona L (2010) A note on mate allocation for dominance handling in genomic selection. *Genet Sel Evol* 42:33. <http://www.gsejournal.org/content/42/1/33>
  75. Shuster DE, Kehrli ME, Ackermann MR, Gilbert RO (1992) Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. *Proc Natl Acad Sci USA* 89:9225–9229
  76. Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, Carta E, Dardano S, Dive M, Fasquelle C, Frennet JC, Hanset R, Hubin X, Jorgensen C, Karim L, Kent M, Harvey K, Pearce BR, Simon P, Tama N, Nie H, Vandeputte S, Lien S, Longeri M, Fredholm M, Harvey RJ, Georges M (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat Genet* 40:449–454

## Antibiotics for Emerging Pathogens

VINAYAK AGARWAL<sup>1</sup>, SATISH K. NAIR<sup>2</sup>

<sup>1</sup>Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

<sup>2</sup>Department of Biochemistry and Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Emerging Targets of Antibiotic Molecules  
 Emerging Chemical Classes of Antibiotic Molecules  
 Future Directions  
 Acknowledgments  
 Bibliography

### Glossary

**Microorganism** A microscopic or submicroscopic organism, too small to be seen by unaided human eye, comprising of bacteria, virus, yeast, protozoa and fungi.

**Pathogen** A disease-causing microorganism which may or may not be infectious.

**Antibiotic** A chemical substance, usually organic in nature, capable of destroying or inhibiting the growth of pathogenic microorganisms.

**Antibiotic resistance** The developed or acquired ability of antibiotic susceptible pathogenic microorganisms to grow and survive despite the inhibitory action of the antibiotic molecules.

**Plasmid** Small, linear or circular genetic elements which can replicate independently of the chromosomal DNA inside a cell.

**Tuberculosis** A highly contagious bacterial disease of humans and animals caused by various strains of *Mycobacterium*, which normally affects the lungs but can also spread to other organs of the body.

**Malaria** A highly contagious protozoal disease caused by various strains of mosquito-borne *Plasmodium*.

**Peptide** Short chains of amino acids connected by peptide bonds.

**Polyketide** Compounds characterized by more than two carbonyl groups connected by single intervening carbon atoms.

## Definition of the Subject

Antibiotics are organic small molecules, many of which are from natural sources, which are used to treat human infections caused by pathogenic microorganisms. While most validated antibiotics are initially very useful clinically, the pathogenic microorganisms that these compounds target are able to evade the action of the antibiotic by development of resistance mechanisms, which eventually render these antibiotics ineffective. Moreover, these resistance mechanisms can be passed on among different types of bacteria in a very simplistic manner that further compromises the usefulness of antibiotics. As a consequence, many diseases that were thought to have been eradicated by antibiotics (such as tuberculosis) have reemerged within these antibiotic-resistant strains. Hence, there is a constant need for the development of new and better antibiotic molecules that can be used to target these drug-resistant microbial populations.

Traditionally, the discovery and development of new antibiotics and other drug molecules has relied

on two parallel and complimentary approaches: the discovery of new molecules of clinical relevance from natural sources, and the use of synthetic chemistry methods to derive compounds, based on the scaffold of naturally occurring molecules, with enhanced favorable antimicrobial properties. More recent development includes the identification of molecules without any natural precedents, based on the use of high-throughput screening methods of potential antibiotic targets against a large number of synthetic compound libraries to identify new classes of antibiotic molecules. Continued developments in all of these approaches, by an amalgam of both academic and industrial efforts, are essential for the development of compounds aimed at treating the evolution of drug-resistant pathogenic microorganisms.

## Introduction

The interaction of prokaryotic microorganisms with human beings is widespread and significant, and the number of bacterial cells within the human body far exceeds the number of the human eukaryotic cells. This interaction between humans and bacteria is designed to be mutually beneficial to both [1, 2]. However, a small number of these bacterial microorganisms are known to cause disease in human beings and are hence called pathogens. The mechanisms by which these pathogenic organisms cause disease are extremely diverse and affect nearly all types of tissues and organs. The advancement of medical science has resulted in the development of two frontline defense mechanisms to fight infections and diseases caused by microbes. The first line of defense is the vaccines that specifically elicit the human host's adaptive immune response to recognize and eliminate pathogenic and harmful microbes, while selectively preserving the beneficial microbe population. Vaccines are generally biological macromolecules, derived from the pathogenic microbes themselves, which provide the adaptive immune response with the information necessary to elicit an antibody or cellular immune response against future invasions by the pathogenic microbes. Vaccines have been developed against several pathogens and have led to very significant reductions of deadly diseases such as tuberculosis and tetanus. However, vaccinations are not currently available against all pathogens, and the

natural evolution of the pathogens themselves may render these vaccinations ineffective against future infections, as exemplified by the inherent difficulties in vaccine development against viral infections such as influenza and HIV/AIDS.

The second frontline defense against pathogenic infections is the discovery and development of small molecule antibiotics that are designed to target and kill the harmful microorganisms with minimal side effects against the human host. Antibiotics are naturally occurring molecules produced by a number of different bacterial species to give them a selective competitive advantage for nutrients or other growth necessities over other microorganisms with which they share their surroundings. Serendipitous discovery and exploitation of these molecules dates back to ancient history [3–5] with numerous early reports, such as the use of the bark from cinchona tree, which contains the antibiotic quinine, used to treat malaria. Antibiotics generally have well-defined mechanisms of action against microbes and target specific biological processes to either retard growth (bacteriostatic antibiotics) or kill the microbe (bacteriocidal antibiotics).

The first report of an isolated antibiotic molecule used to effectively treat infections was the development and use of penicillin from the fungus *Penicillium notatum* by Alexander Fleming in 1928. Penicillin was used to treat diseases such as syphilis and other infections caused by staphylococci and streptococci. The term “antibiotic” was coined by the microbiologist Selman Waksman in 1942 [6] and is derived from the Greek *antibiosis* (“against life”). Hailed as a wonder drug, the discovery of penicillin marked the beginning of the antibiotic age, during which time antibiotics were developed to treat almost all major human infectious diseases. During this era, thousands of antibiotic molecules were discovered and these compounds were applied to fight bacterial, fungal, protozoan, and yeast infections. Concurrently, synthetic organic chemistry was also applied for the production of derivatives of naturally occurring antibiotics, and these synthetic compounds exhibited more desirable properties, such as decreased incidences of allergic reactions in humans. The discovery and development of these molecules marked a decreased incidence of outbreaks of the most common infectious diseases

and contributed to an increase in life expectancy around the world.

However, repeated challenge of the pathogenic microorganisms with these antibiotics, in combination with the natural evolution of their genetic information, has resulted in the development of antibiotic resistance in these pathogens [7–9]. This process has led to the development of microbial resistance against nearly all available antibiotic molecules and has precipitated to an immense medical and financial challenge. The arsenal available to fight infections is becoming increasingly limited, and the rate of development of new molecules has not kept pace with the emergence of antibiotic resistance.

### Factors Contributing to Increased Antibiotic Resistance

1. Widespread and indiscriminate overuse of antibiotics in the clinical environment, which exposes pathogenic microorganisms to an ever increasing concentration of different antibiotic molecules, which in turn increases the selection pressure to develop resistance against these antibiotic molecules [10, 11].
2. Use of antibiotics in agriculture and livestock rearing also leads to increased exposure of pathogenic bacteria to antibiotic molecules. Many of the antibiotics in use in agriculture and livestock feeds are also used for treating humans, and hence, development of resistant bacterial strains can be transferred from one setting to the other very easily [12–14].
3. Genetic recombination and transfer of genes from an antibiotic-resistant microbe to a non-resistant microbe, which leads to spread of resistance among microbial species. Antibiotic resistance genes are usually borne on highly mobile genetic vehicles called plasmids, which can easily be transmitted among microbes [15]. Microorganisms are able to sequester multiple plasmids, resulting in the evolution of plasmids that contain more than one resistance determinant. For example, the *Shigella* epidemic that caused nearly a quarter million deaths in Guatemala in 1968 was caused by a pathogen that contained a plasmid with resistance against four of the most commonly used antibiotics [16–18].

## Mechanisms by Which Microorganisms Develop Resistance to Antibiotics

1. Efflux pumps in the microbial cell walls and membranes which actively transport the antibiotic molecule out of the cell and hence decrease its concentration within the cell. Consequently, increasing doses of antibiotics are required to target such microbes [19–21]. This mechanism is typified by the resistance to the broad-spectrum antibiotic tetracycline, which is mediated by the genetically mobile *tet* genes that encode for active efflux pumps for tetracycline and its derivatives [22–24].
2. Chemical modification of the antibiotic molecule by specific enzymes within the target pathogenic microbe, which render the antibiotic ineffective [25]. This mechanism is of particular concern against  $\beta$ -lactam antibiotics, such as penicillin and its derivatives [26, 27].
3. Modulation of the antibiotic target within the target cell so that the antibiotic is no longer able to bind and engage the target. Random mutations within the antibiotic target genes may lead to the emergence and subsequent selection of microbe species with a desensitized antimicrobial target. This mechanism of resistance is operative for development of resistance against the macrolide antibiotic erythromycin, and resistance results from the modification of its target, the bacterial ribosome, such that the antibiotic is no longer able to find the target [28, 29]. Another example is resistance to the “drug of last resort” vancomycin, which results as a change in its target of the D-Ala- D-Ala dipeptide moiety of the bacterial cell wall to D-Ala- D-Lac or to D-Ala- D-Ser [30–32].
4. Alteration of the metabolic pathways that are targeted by the antibiotic molecule so that inhibition of these metabolic pathways is not inhibitory or lethal to the microbe. A primary example of this class of resistance is the alterations in the sterol biosynthesis pathway in fungi which confer resistance to azole antibiotics [33, 34].

In light of the above factors, there is an ever-growing need for academic laboratories as well as for the pharmaceutical industries to search for antibiotics that have new and as yet unexplored mode of actions. New chemical classes of drug molecules are needed to

mitigate the effects of resistance against the current drug chemical entities.

## Emerging Targets of Antibiotic Molecules

Several new classes of antimicrobial targets have been identified, and drug development based on these findings, though still in its infancy, have taken major leaps forward.

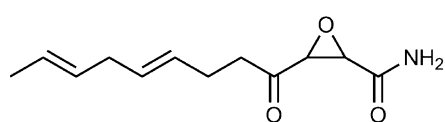
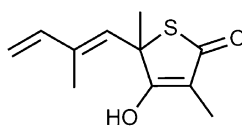
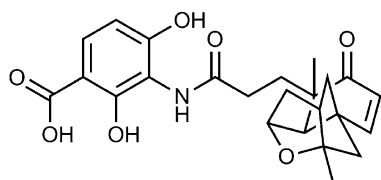
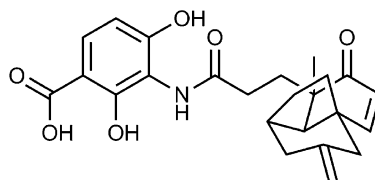
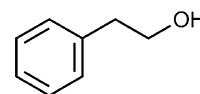
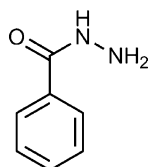
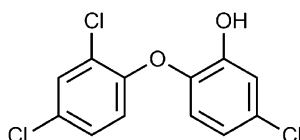
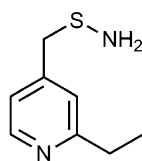
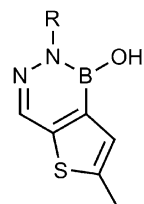
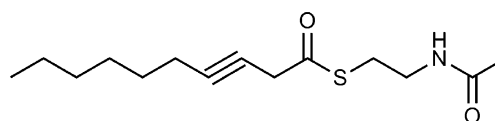
### Microbial Fatty Acid Biosynthesis Inhibition

The microbial fatty acid biosynthesis pathway presents numerous targets for drug development, particularly against the causative agent of tuberculosis – *Mycobacterium tuberculosis*. The metabolic pathways and chemical mechanisms for the synthesis of fatty acids are shared between a majority of prokaryotes and eukaryotes, including humans. However, as a result of differences in protein sequence and different arrangement of the active sites, compounds that target the bacterial fatty acid synthesis pathway are not cross-reactive or toxic against humans. The most well-characterized class of fatty acid biosynthesis inhibitor molecules is the antituberculosis molecule isoniazid, which inhibits the biosynthesis of mycolic acid, an essential component of the mycobacterial cell wall. Triclosan, an extensively used molecule in consumer products, such as toothpastes and mouthwashes, is also a fatty acid biosynthesis inhibitor.

An example of an emerging molecule in this class of antimicrobials is platensimycin (Fig. 1), produced by the organism *Streptomyces platensis* [35], which inhibits the growth of several pathogenic Gram-positive bacteria such as *Staphylococcus aureus*, *Enterococcus faecalis*, and *Staphylococcus pneumonia* [36]. Platensimycin inhibits the 3-ketoacyl-ACP synthase (KAS) enzyme FabF. Continued high-throughput screening and host genetic engineering efforts led to the identification of a platensimycin analog, platencin [37–39], which, in addition to FabF, inhibits another KAS enzyme called FabH and displays enhanced pharmacodynamic properties as compared to platensimycin. Chemical structures of microbial fatty acid biosynthesis inhibitor molecules discussed in this section are illustrated in Fig. 1.

Other KAS inhibitor molecules include cerulenin and thiolactomycin [40, 41]. Cerulenin is a



**KAS Inhibitors****CERULENIN****THIOLACTOMYCIN****PLATENSIMYCIN****PLATENCIN****PlsB Inhibitors****PHENETHYL ALCOHOL****FabI Inhibitors****ISONIAZID****TRICLOSAN****ETHIONAMIDE****DIAZOBORINES****FabA Inhibitors****3-DECYNOYL-N-ACETYLCYSTEAMINE****Antibiotics for Emerging Pathogens. Figure 1**

Chemical structures of representative microbial fatty acid inhibitor molecules

broad-spectrum antifungal antibiotic produced by fungus *Caephalosporium caerulens* and preferentially inhibits the KAS enzymes FabB and FabF. Cerulenin forms a stable covalent adduct with the active site cysteine residue of these enzymes, and thus inhibits malonyl-group incorporation in the growing fatty acid chains [40, 42]. Thiolactomycin and related compounds have been isolated from various *Streptomyces* and show potent inhibitory activity against fatty acid biosynthesis and mycolic acid biosynthesis in mycobacteria as well as against protozoan parasites such as malaria-causing *Plasmodium falciparum*. Thiolactomycin mimics the binding of malonyl coenzyme A molecule in the active site of KAS enzymes

FabB, FabF, and FabH [41, 43]. Attempts for the preparation of semisynthetic derivatives of thiolactomycin are currently underway, and some of these derivatives are expected to improve upon its antimicrobial properties of thiolactomycin [44].

Molecules such as triclosan, isoniazid, diazaborines, and ethionamide target the enoyl-ACP reductase enzyme FabI of the fatty acid biosynthesis pathway. Each of these molecules binds in the active site of the enzyme and forms a tight complex with the nicotinamide adenine dinucleotide (NAD<sup>+</sup>) cofactor required for catalysis by FabI. However, pseudomonads and *S. pneumoniae* contain an additional enoyl-ACP reductase enzyme, FabK, which is not susceptible to

inhibition by these compounds, which allows these pathogens to overcome the antibacterial activities of these compounds.

The fatty acid biosynthesis pathways also present several other enzymes that can be targeted for drug development. The isomerase activity of the enzyme FabA can be inhibited by the drug candidate compound 3-decynoyl-*N*-acetylcysteamine. Metabolic interaction of the acetyl coenzyme A carboxylase enzyme with the fatty acid biosynthetic pathway also poses a potential target for inhibitor screening. Finally, compounds such as phenethyl alcohol, which inhibit the enzyme PlsB, responsible for the link between fatty acid biosynthesis and phospholipid biosynthesis, are also antimicrobial agents.

### Isoprenoid Biosynthesis Inhibition

Isoprenoids are naturally occurring diverse organic compounds derived from the condensation of two or more isoprene monomeric units. They represent the largest class of small molecules on Earth. The isoprenoid pathways between prokaryotes and eukaryotes are very divergent: Humans synthesize isoprenoids by the mevalonate pathway, while microbes synthesize isoprenoids via the nonmevalonate pathways. Both pathways present numerous potential drug targets. One of the most widely used cholesterol-lowering drug, Lipitor, targets the mevalonate pathway in humans which is responsible for cholesterol biosynthesis. Antimicrobials such as terbinafine target the nonmevalonate pathways in fungi and yeast.

Several promising lead compounds that target the nonmevalonate pathway in the malaria parasite *Plasmodium falciparum* have been characterized. Principal among these are the phosphonate molecules fosmidomycin, first isolated from *Streptomyces lavendulae*, and its derivative FR900098, isolated from *Streptomyces rubellomurinus* [45, 46]. Fosmidomycin inhibits the enzyme 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR) of the malaria parasite and has progressed to the stage of human phase II clinical trials for use in combination therapy with lincosamide antibiotic clindamycin [47–49]. However, the bioavailability of this molecule is low in bacterial models, with tuberculosis-causing bacterial pathogen essentially resistant to this molecule due to lack of uptake [50].

FR900098 is a structural analog of fosmidomycin and has been shown to demonstrate better efficacy in vivo than the parent compound. Synthetic and biosynthetic routes for the production and manipulation of this compound are currently being pursued [51, 52], as is the synthesis of several other structural analogs of fosmidomycin [53]. Fosmidomycin and FR900098 belong to the phosphonate class of antimicrobials, which are characterized by the presence of a stable carbon–phosphorous covalent linkage [54].

In addition to the DXR enzyme, a second potential drug target is the isoprenoid biosynthesis enzyme IspH, which is present in the vast majority of pathogenic bacteria and the malaria parasite but not in humans [55]. IspH is an attractive target for the development of inhibitors which could be potential drug and herbicide candidates. A series of diphosphate and bisphosphonate compounds have been synthesized, some of which inhibited the IspH enzyme at nanomolar concentrations [56, 57].

Other isoprenoid biosynthetic drug targets include the enzyme dehydrosqualene synthase CrtM, from the human pathogen *Staphylococcus aureus*, which can be inhibited by phosphonosulfonates and phosphonoacetamide compounds [58]. These compounds lead to inhibition of production of the *S. aureus* virulence factor staphyloxanthin, which makes these otherwise recalcitrant pathogenic cells susceptible to reactive oxygen species and subsequent clearance by the human immune system [59, 60]. Finally, the antiarrhythmia drug amiodarone, which blocks various metal ion channels in the cell membrane [61], in conjunction with itraconazole, a fungal cytochrome P450 inhibitor [62], is also being reevaluated for the treatment of Chagas disease and cutaneous leishmaniasis [63].

### Cell Wall Biosynthesis Inhibition

Inhibition of cell wall biosynthesis has been one of the earliest validated targets for antibiotic development and still remains one of the most fruitful areas for the development of new drugs. However, drug development against cell wall biosynthesis is particularly challenging due to the presence of an additional lipopolysaccharide layer in Gram-negative bacteria and the fungal chitin cell wall in fungi, both of which are extremely rigid and impermeable to most small

molecule drugs. The following section focuses on some new nucleosidic antibiotic molecules that inhibit microbial cell wall biosynthesis and hold promise for future development. Chemical structures of microbial cell wall biosynthesis inhibitor molecules discussed in this section are illustrated in Fig. 2.

Nucleoside antifungal polyoxin molecules, such as nikkomycins, were isolated from several strains of *Streptomyces* [64] and have been shown to inhibit the biosynthesis of the fungal chitin cell wall. Nikkomycins display a wide spectrum of bioactivity while being nontoxic to animals and plants and have traditionally been used for controlling plant fungal diseases in agriculture. The parent nikkomycin compounds show relatively weak activity against opportunistic human pathogenic fungi, and semisynthetic derivatives are now being prepared which may display better pharmacological properties [65].

Capuramycin was first isolated from *Streptomyces griseus* in the 1980s and shows potent antibacterial properties against the human pathogens *Streptococcus pneumoniae* and *Mycobacterium smegmatis* [66]. Structure activity relationships have been explored by the synthesis of capuramycin derivatives [67, 68].

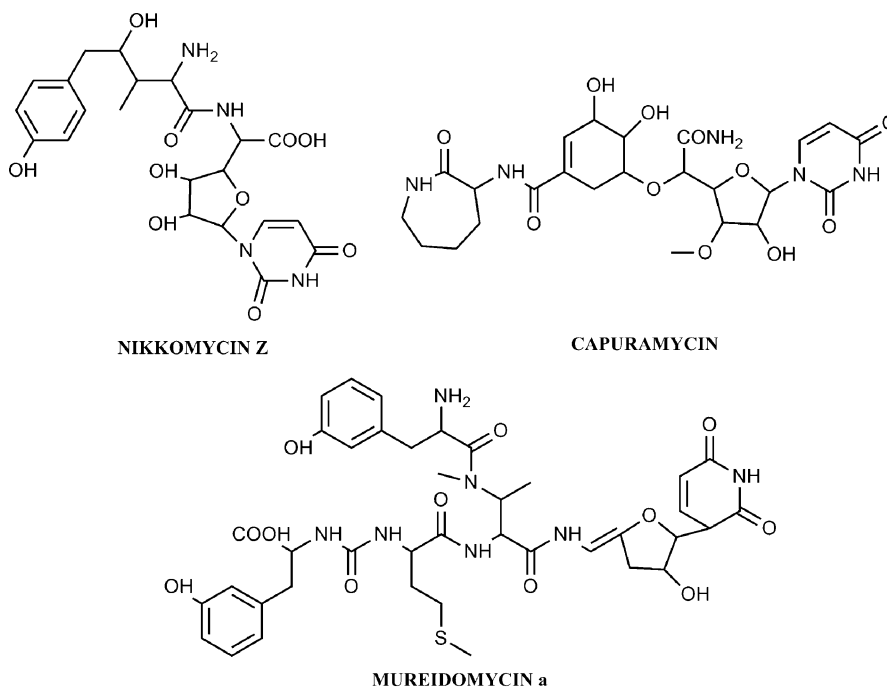
In *in vivo* studies against *Mycobacterium tuberculosis*, capuramycin demonstrated unique phenotypic changes in the pathogen which are not shown by any other currently available antibiotic and which may signify a mode of action distinct from those of known antibiotics [69].

Structurally related uridylpeptide molecules such as mureidomycins, pacidamycins, napsamycins, and sansanmycins have demonstrated activity against infections of the opportunistic human pathogen *Pseudomonas aeruginosa* in rodent models. Such infections are difficult to treat in immune-compromised patients as this pathogen can form a protective biofilm layer resulting in chronic infections [70–72]. These molecules do not show any cross-reactivity with mammalian cells.

## Emerging Chemical Classes of Antibiotic Molecules

### Peptidic Antibiotics

Peptidic antimicrobials are typically synthesized by the ribosome and consist of short polymers of the 20 naturally occurring amino acids linked together by



**Antibiotics for Emerging Pathogens. Figure 2**

Chemical structures of representative microbial cell wall biosynthesis inhibitor molecules

amide bonds. A second class of peptide antimicrobials are generated by a molecular assembly line, called nonribosomal peptide synthases (NRPS), which can incorporate nonnatural amino acids. The use of peptides to target pathogenic microbes is exemplified by the production of short cationic peptides by the human and plant immune systems, which either recruit the immune system for pathogen clearance or disrupt the outer cell and organelle cell membranes of microbes. A common conserved feature among all peptidic antimicrobials is posttranslational modifications conferred upon the peptide backbone, which are essential for their biological activities [73–75].

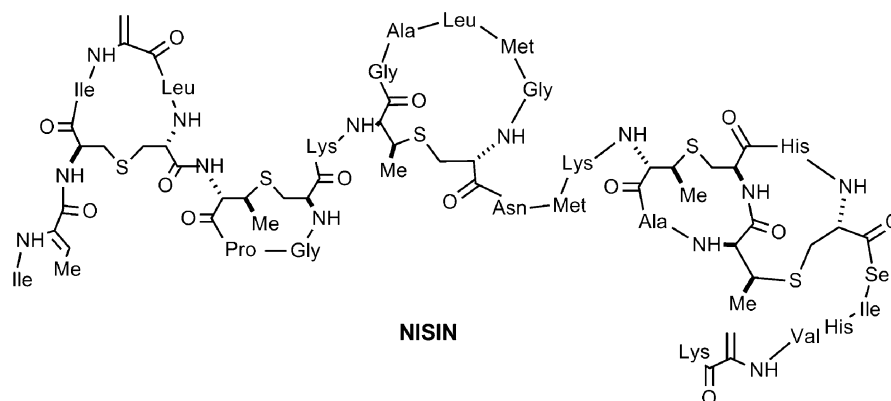
A well-known example of one class of ribosomally produced antimicrobial peptides is lantibiotics. Although these molecules are synthesized on the ribosome, they undergo posttranslational processing, resulting in the formation of several ring-like structures that add rigidity [76]. These compounds can be further modified on the amino and carboxy termini of these peptides as well, as shown in the structures of the lantibiotics epilancin [77] and microbisporicin [78]. Lantibiotics demonstrated multiple modes of action and exert their biological activities by binding to the lipid II cell wall moiety in bacteria [79], which leads to both inhibition of peptidoglycan biosynthesis as well as cell membrane permeabilization.

The most extensively used and well-characterized lantibiotic to date is Nisin (Fig. 3), initially identified as a by-product of milk fermentation, which contains seven lantibiotic rings, and has been used in the livestock industry for the treatment of bovine mastitis

[80, 81]. Nisin, produced by the bacterium *Lactococcus lactis*, shows no adverse effects in humans and is being evaluated in human trials for the treatment of staphylococcal mastitis in lactating women [82].

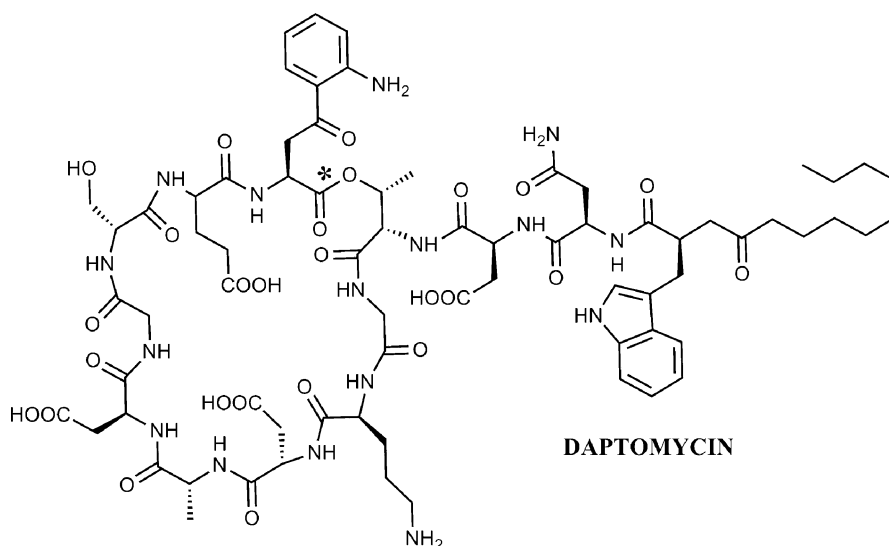
Among the NRPS peptides, the most promising drug candidates are the lipopeptide antibiotic daptomycin and glycopeptides vancomycin and teicoplanin. Lipopeptide antibiotic such as daptomycin (Fig. 4) contains a long lipid chain attached to the peptide, which helps in targeting the bacterial cell wall. Daptomycin is produced by *Streptomyces roseosporus*, and genetic manipulations of the producer organism have led to the production of a combinatorial library of daptomycin derivatives [83–86]. Daptomycin shows potent antimicrobial activity against methicillin-resistant *Staphylococcus aureus* (MRSA) [87], infections by which are a rising global concern, and is already registered for the treatment of skin and soft tissue infections [88], bacteremia, and bacterial endocarditis [89].

The glycopeptide antibiotics are produced as cyclic aglycones (without sugars) and are then decorated by a number of different sugar molecules that are covalently attached by dedicated glycosyltransferase enzymes [90–92]. Among these, vancomycin (Fig. 5) [93], which prevents cell wall biosynthesis in Gram-positive bacteria, is clinically approved as the “drug of last resort” against MRSA infections [94]. Oritavancin (Fig. 5), a semisynthetic derivative of vancomycin, is potent against vancomycin-resistant enterococci (VRE) and is currently under human trials for the treatment of soft tissue infections by Gram-positive



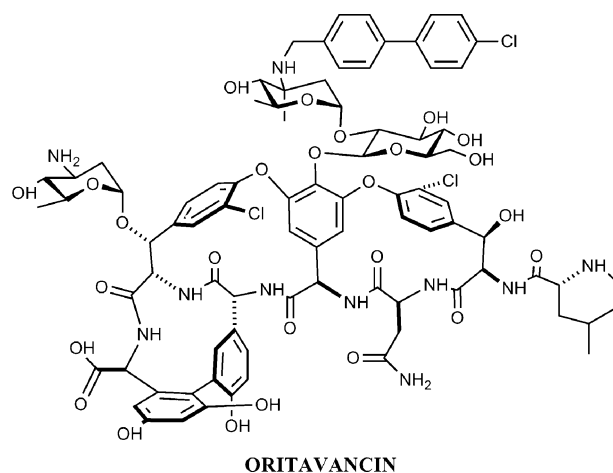
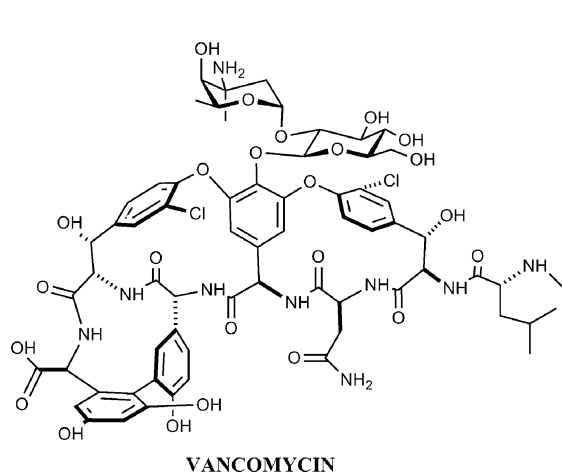
**Antibiotics for Emerging Pathogens. Figure 3**

Chemical structures of lantibiotic nisin. Some amino acids are abbreviated by their three-letter codes



**Antibiotics for Emerging Pathogens. Figure 4**

Chemical structures of lipopeptide antibiotic daptomycin



**Antibiotics for Emerging Pathogens. Figure 5**

Chemical structures of representative glycopeptides and lipoglycopeptide antibiotics vancomycin and oritavancin

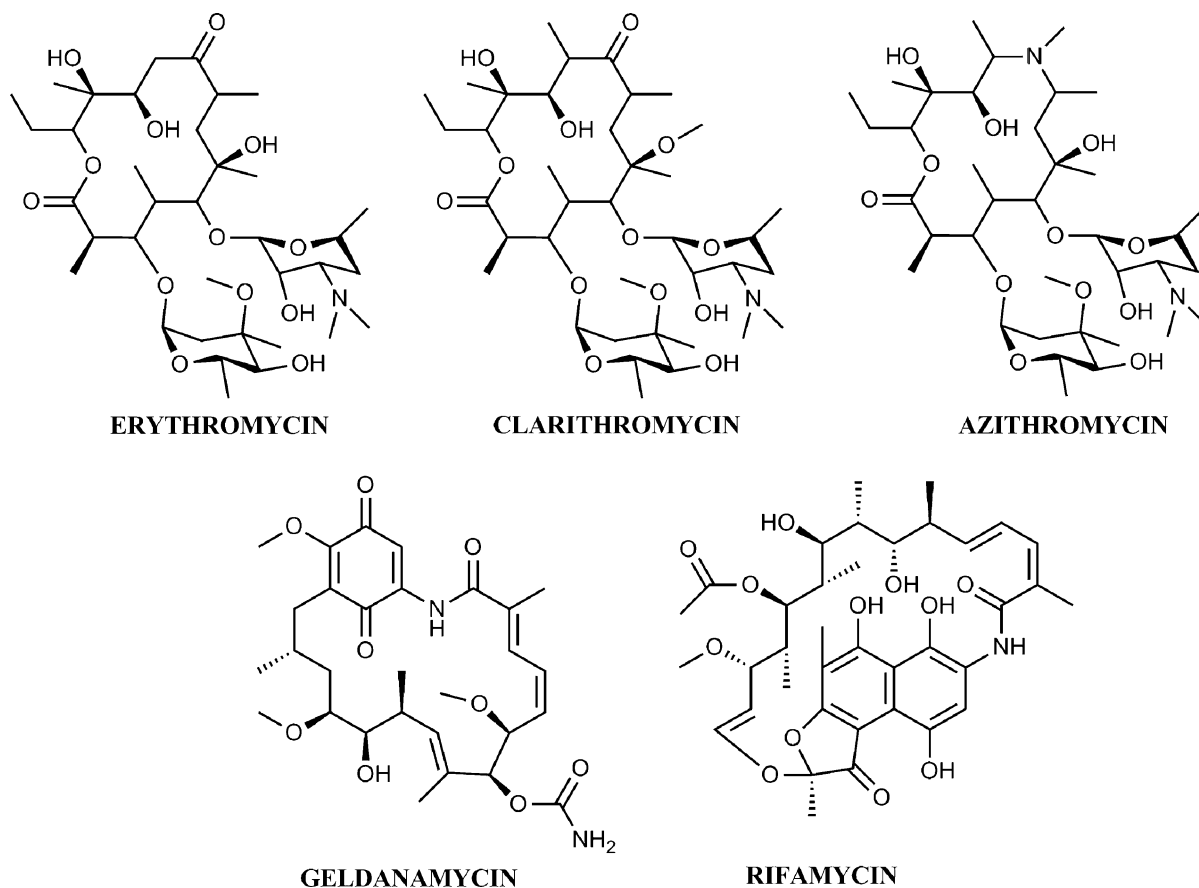
bacteria [95]. Also of note are the cyclic lactone linkage-containing depsipeptides katanosins/plusbacins [96] and glycolipodepsipeptide ramoplanin [97], which are also active against MRSA.

### Polyketide Antibiotics

Polyketide antibiotics are similar to nonribosomal peptide antibiotics as they are synthesized by

multimodular polyketide synthases. These enzymes are analogous to the nonribosomal peptide synthases but typically catalyze the condensation of malonyl coenzyme A-derived monomeric units into cyclic lactone rings, onto which further modifications such as hydroxylations, glycosylations, and methylations are added. The most well-known polyketide antibiotic is erythromycin (Fig. 6), produced by *Streptomyces erythreus*, which has been in clinical use for more





**Antibiotics for Emerging Pathogens. Figure 6**  
Chemical structures of representative polyketide antibiotics

than 50 years. Several derivatives of erythromycin, such as clarithromycin and azithromycin (Fig. 6), have been developed for clinical use [98]. The mode of action of these compounds is inhibition of protein synthesis through binding to the various subunits of the bacterial ribosome [99].

Advancements in genetic technologies have opened up avenues for the manipulation and reprogramming of polyketide synthases for the production of new and novel molecules, which may yield drug candidates with enhanced pharmacological properties [100–102]. Of particular note are the ansamycin group of macrolide molecules, which differ from the molecules previously discussed in that the monomeric unit is 3-amino-5-hydroxybenzoic acid rather than a malonyl coenzyme A derivative. Various ansamycin molecules, such as rifamycin, geldanamycin, herbimycin, and

ansamitocins, have been isolated, and their biosynthetic routes have been characterized. Of these, rifamycin, which inhibits bacterial RNA polymerase [103], is widely used in the treatment of tuberculosis. Geldanamycin, biosynthesized by *Streptomyces hygroscopicus*, and its analogs [104] inhibit heat shock protein 90 (Hsp90) [105] and are being evaluated in human clinical trials as antitumor agents [106].

### Trojan Horse Antibiotics

A limiting factor for many antimicrobials and antibiotics is the lack of permeability across the microbial cell membranes. The outer membrane of the Gram-negative bacteria presents a formidable obstacle for drug entry [107]. Such Gram-negative bacteria employ a variety of active importer channels and pores to

internalize requisite nutrient material. Nature has ingeniously designed several antimicrobial compounds that utilize a “Trojan horse” strategy to overcome permeability barriers; specifically, these drug molecules are covalently attached to a carrier that facilitates recognition and import through the active transporters. Once inside the target cell, the carrier is removed to release the active drug, which then exerts its antimicrobial activities. This strategy for drug delivery has several advantages. First, active transport of the molecule inside susceptible cells alleviates the need for very high doses for the drug. Second, targeting of the molecule is highly specific only for those microbial cells that have the required transporter to internalize the molecules. Consequently, these “Trojan horse” are not taken up in humans and can be effective microbials even at very low doses.

Notable among this class of compounds are the sideromycins [108, 109], which are characterized by the attachment of the antibiotic molecule to ferric ion-binding chemical components called siderophores. The siderophore-iron complexes are specifically recognized and internalized by dedicated transport machinery in bacteria. Sideromycins consists of structurally diverse molecules including danomycins, salmycins, albomycins, ferrimycins, and microcins. Bioactivities for all these molecules have been established against several human pathogens [110]. Interestingly, as rapidly dividing cancerous cells also display siderophore uptake receptors on their cell surfaces, these molecules are also being evaluated as potential anticancer compounds [109]. Several synthetic conjugates of antibiotic molecules with siderophores have also been generated for evaluation of antimicrobial properties [111].

A different class of “Trojan horse” molecules consists of drugs that are attached to peptide carrier molecules. An example of one such drug is the compound (Z)-l-2-amino-5-phosphono-3-pentenoic acid (APPA), which is a potent inhibitor of the metabolic enzyme threonine synthase, and the “Trojan horse” version consists of APPA attached to a dipeptide. Depending on the amino acids in the dipeptide, the resultant conjugate can either be an antifungal (rhizotocins) or antibacterial (plumbemycin) [112]. Thus, a variety of activities can be encoded in the identical molecule by simply altering the nature of the conjugated peptide. A similar strategy

for peptide attachment is utilized in the protein synthesis inhibitor microcin C7 [113].

### Future Directions

Maintenance of global health in light of the reemergence of life-threatening infectious diseases requires a continuous influx of new and potent antibiotic molecules that are efficacious and safe for human and animal clinical use. Pathogenic organisms have displayed a remarkable resilience in both developing and propagating mechanisms for overcoming the activities of these antibiotic molecules. The development of new antibiotic molecules relies on two principal paradigms: the characterization of new microbial targets against which new drug molecules can be developed, and the identification of new chemical entities which can be used as antibiotics. Both routes have resulted in the production of very successful drug molecules, and future prospects for antibiotic development are reliant on further progress in these directions. Recent biotechnological advances in the area of microbial genome sequencing have led to a rapid increase in exploration of the microbial metabolome for both novel pathways to target in pathogens and new candidates for use as antibiotics. Semisynthetic chemical derivatization of naturally occurring antibiotics has further expanded upon the repertoire of candidate drugs, and combinatorial synthetic approaches have generated vast chemical libraries which may be screened for compounds with desired biological properties.

Future antibiotic development will require an amalgamation of various academic and industrial approaches such as (1) systems biology to explore metabolic connections and effects of antibiotic molecules on the human host and targeted pathogenic microbes, (2) synthetic biochemistry to generate and diversify arrays of compounds, (3) genomics and bioinformatics to identify and analyze antibiotic biosynthetic gene clusters, and (4) microbial ecological evaluations to better understand the role of naturally occurring antibiotic scaffolds in their native microbial niches. Another promising developing area is combinatorial therapy, in which multiple antibiotic molecules are administered together in order to overcome microbial resistance mechanisms and achieve better efficacy and pathogen clearance. This approach has

proven to be very successful in antiviral and oncological applications, and combination regimens have been used for the treatment of tuberculosis for several decades now. Another area of rapid progress in the future promises to be the development of narrow-spectrum antibiotics and better diagnostic tools which would aid in curtailing the indiscriminate use of broad-spectrum antibiotics, and thus reduce the incidence of microbial resistance against antibiotics.

### Acknowledgments

Research in the Nair lab is supported by the NIGMS. We thank Neha Garg, Yue Hao, and Zhi Li for stimulating discussions.

### Bibliography

- Sears CL (2005) A dynamic partnership: celebrating our gut flora. *Anaerobe* 11(5):247–251
- Backhed F et al (2005) Host-bacterial mutualism in the human intestine. *Science* 307(5717):1915–1920
- Lindblad WJ (2008) Considerations for determining if a natural product is an effective wound-healing agent. *Int J Low Extrem Wounds* 7(2):75–81
- Forrest RD (1982) Early history of wound treatment. *J R Soc Med* 75(3):198–205
- Forrest RD (1982) Development of wound therapy from the dark ages to the present. *J R Soc Med* 75(4):268–273
- Waksman SA (1947) What is an antibiotic or an antibiotic substance? *Mycologia* 39(5):565–569
- Davies J, Davies D (2010) Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* 74(3):417–433
- Allen HK et al (2010) Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol* 8(4):251–259
- Moellering RC Jr (1998) Antibiotic resistance: lessons for the future. *Clin Infect Dis* 27(Suppl 1):S135–S140 discussion S141–2
- Goossens H et al (2005) Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet* 365(9459):579–587
- Tacconelli E et al (2008) Does antibiotic exposure increase the risk of methicillin-resistant *Staphylococcus aureus* (MRSA) isolation? A systematic review and meta-analysis. *J Antimicrob Chemother* 61(1):26–38
- Mathew AG, Cissell R, Liamthong S (2007) Antibiotic resistance in bacteria associated with food animals: a United States perspective of livestock production. *Foodborne Pathog Dis* 4(2):115–133
- Angulo FJ, Nargund VN, Chiller TC (2004) Evidence of an association between use of anti-microbial agents in food animals and anti-microbial resistance among bacteria isolated from humans and the human health consequences of such resistance. *J Vet Med B Infect Dis Vet Public Health* 51(8–9):374–379
- Singer RS et al (2003) Antibiotic resistance—the interplay between antibiotic use in animals and human beings. *Lancet Infect Dis* 3(1):47–51
- Koch AL (1981) Evolution of antibiotic resistance gene function. *Microbiol Rev* 45(2):355–378
- Mendoza MC et al (1988) Plasmid typing of *Shigella sonnei* epidemic strains and molecular relationship of their R-plasmids. *Eur J Epidemiol* 4(2):158–163
- Yagupsky P et al (1991) Use of multiple markers for investigation of an epidemic of *Shigella sonnei* infections in Monroe County, New York. *J Clin Microbiol* 29(12):2850–2855
- Kariuki S et al (1996) Molecular typing of multi-drug resistant *Shigella dysenteriae* type 1 by plasmid analysis and pulsed-field gel electrophoresis. *Trans R Soc Trop Med Hyg* 90(6):712–714
- Li XZ, Nikaido H (2009) Efflux-mediated drug resistance in bacteria: an update. *Drugs* 69(12):1555–1623
- Lin L, Ling BD, Li XZ (2009) Distribution of the multidrug efflux pump genes, *adeABC*, *adeDE* and *adeIJK*, and class 1 integron genes in multiple-antimicrobial-resistant clinical isolates of *Acinetobacter baumannii*-*Acinetobacter calcoaceticus* complex. *Int J Antimicrob Agents* 33(1):27–32
- Van Bambeke F, Balzi E, Tulkens PM (2000) Antibiotic efflux pumps. *Biochem Pharmacol* 60(4):457–470
- Chopra I, Roberts M (2001) Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol Mol Biol Rev* 65(2):232–260, second page, table of contents
- Rodrigues L et al (2009) The role of efflux pumps in macrolide resistance in *Mycobacterium avium* complex. *Int J Antimicrob Agents* 34(6):529–533
- Pages JM, Amaral L (2009) Mechanisms of drug efflux and strategies to combat them: challenging the efflux pump of gram-negative bacteria. *Biochim Biophys Acta* 1794(5):826–833
- Wright GD (2005) Bacterial resistance to antibiotics: enzymatic degradation and modification. *Adv Drug Deliv Rev* 57(10):1451–1470
- Bradford PA (2001) Extended-spectrum beta-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat. *Clin Microbiol Rev* 14(4):933–951
- Jacoby GA, Sutton L (1985) Beta-Lactamases and beta-lactam resistance in *Escherichia coli*. *Antimicrob Agents Chemother* 28(5):703–705
- Lovmar M et al (2009) Erythromycin resistance by L4/L22 mutations and resistance masking by drug efflux pump deficiency. *EMBO J* 28(6):736–744
- Walsh C (2000) Molecular mechanisms that confer antibacterial drug resistance. *Nature* 406(6797):775–781
- Bugg TD et al (1991) Molecular basis for vancomycin resistance in *Enterococcus faecium* BM4147: biosynthesis of a depsipeptide peptidoglycan precursor by vancomycin

- resistance proteins VanH and VanA. *Biochemistry* 30(43):10408–10415
31. Bugg TD et al (1991) Identification of vancomycin resistance protein VanA as a D-alanine:D-alanine ligase of altered substrate specificity. *Biochemistry* 30(8):2017–2021
  32. Walsh CT et al (1996) Bacterial resistance to vancomycin: five genes and one missing hydrogen bond tell the story. *Chem Biol* 3(1):21–28
  33. Joseph-Horne T, Hollomon DW (1997) Molecular mechanisms of azole resistance in fungi. *FEMS Microbiol Lett* 149(2): 141–149
  34. Lupetti A et al (2002) Molecular basis of resistance to azole antifungals. *Trends Mol Med* 8(2):76–81
  35. Herath KB, Attygalle AB, Singh SB (2007) Biosynthetic studies of platensimycin. *J Am Chem Soc* 129(50):15422–15423
  36. Wang J et al (2006) Platensimycin is a selective FabF inhibitor with potent antibiotic properties. *Nature* 441(7091): 358–361
  37. Smanski MJ et al (2009) Engineered *Streptomyces platensis* strains that overproduce antibiotics platensimycin and platencin. *Antimicrob Agents Chemother* 53(4):1299–1304
  38. Jayasuriya H et al (2007) Isolation and structure of platencin: a FabH and FabF dual inhibitor with potent broad-spectrum antibiotic activity. *Angew Chem Int Ed Engl* 46(25):4684–4688
  39. Wang J et al (2007) Discovery of platencin, a dual FabF and FabH inhibitor with in vivo antibiotic properties. *Proc Natl Acad Sci USA* 104(18):7612–7616
  40. Siggaard-Andersen M et al (1994) The fabJ-encoded beta-ketoacyl-[acyl carrier protein] synthase IV from *Escherichia coli* is sensitive to cerulenin and specific for short-chain substrates. *Proc Natl Acad Sci USA* 91(23):11027–11031
  41. Machutta CA et al (2010) Slow onset inhibition of bacterial beta-ketoacyl-acyl carrier protein synthases by thiolactomycin. *J Biol Chem* 285(9):6161–6169
  42. Johansson P et al (2008) Inhibition of the fungal fatty acid synthase type I multienzyme complex. *Proc Natl Acad Sci USA* 105(35):12803–12808
  43. Nishida I, Kawaguchi A, Yamada M (1986) Effect of thiolactomycin on the individual enzymes of the fatty acid synthase system in *Escherichia coli*. *J Biochem* 99(5):1447–1454
  44. McFadden JM et al (2005) Application of a flexible synthesis of (5R)-thiolactomycin to develop new inhibitors of type I fatty acid synthase. *J Med Chem* 48(4):946–961
  45. Okuhara M et al (1980) Studies on new phosphonic acid antibiotics. III. Isolation and characterization of FR-31564, FR-32863 and FR-33289. *J Antibiot (Tokyo)* 33(1):24–28
  46. Okuhara M et al (1980) Studies on new phosphonic acid antibiotics. I. FR-900098, isolation and characterization. *J Antibiot (Tokyo)* 33(1):13–17
  47. Lell B et al (2003) Fosmidomycin, a novel chemotherapeutic agent for malaria. *Antimicrob Agents Chemother* 47(2):735–738
  48. Missinou MA et al (2002) Fosmidomycin for malaria. *Lancet* 360(9349):1941–1942
  49. Wiesner J, Jomaa H (2007) Isoprenoid biosynthesis of the apicoplast as drug target. *Curr Drug Targets* 8(1):3–13
  50. Brown AC, Parish T (2008) Dxr is essential in *Mycobacterium tuberculosis* and fosmidomycin resistance is due to a lack of uptake. *BMC Microbiol* 8:78
  51. Fokin AA et al (2007) Synthesis of the antimalarial drug FR900098 utilizing the nitroso-ene reaction. *Org Lett* 9(21): 4379–4382
  52. Johannes TW et al (2010) Deciphering the late biosynthetic steps of antimalarial compound FR-900098. *Chem Biol* 17(1):57–64
  53. Kurz T et al (2007) Conformationally restrained aromatic analogues of fosmidomycin and FR900098. *Arch Pharm (Weinheim)* 340(7):339–344
  54. Metcalf WW, van der Donk WA (2009) Biosynthesis of phosphonic and phosphinic acid natural products. *Annu Rev Biochem* 78:65–94
  55. Rohrich RC et al (2005) Reconstitution of an apicoplast-localised electron transfer pathway involved in the isoprenoid biosynthesis of *Plasmodium falciparum*. *FEBS Lett* 579(28):6433–6438
  56. Wang K et al (2010) Inhibition of the Fe(4)S(4)-cluster-containing protein LspH (LytB): electron paramagnetic resonance, metallacycles, and mechanisms. *J Am Chem Soc* 132(19):6719–6727
  57. Wang W et al (2010) Bioorganometallic mechanism of action, and inhibition, of LspH. *Proc Natl Acad Sci USA* 107(10): 4522–4527
  58. Liu CI et al (2008) A cholesterol biosynthesis inhibitor blocks *Staphylococcus aureus* virulence. *Science* 319(5868): 1391–1394
  59. Song Y et al (2009) Inhibition of staphyloxanthin virulence factor biosynthesis in *Staphylococcus aureus*: in vitro, in vivo, and crystallographic results. *J Med Chem* 52(13):3869–3880
  60. Song Y et al (2009) Phosphonosulfonates are potent, selective inhibitors of dehydrosqualene synthase and staphyloxanthin biosynthesis in *Staphylococcus aureus*. *J Med Chem* 52(4):976–988
  61. Kodama I, Kamiya K, Toyama J (1999) Amiodarone: ionic and cellular mechanisms of action of the most promising class III agent. *Am J Cardiol* 84(9A):20R–28R
  62. Isoherranen N et al (2004) Role of itraconazole metabolites in CYP3A4 inhibition. *Drug Metab Dispos* 32(10):1121–1131
  63. Paniz-Mondolfi AE et al (2009) Amiodarone and itraconazole: a rational therapeutic approach for the treatment of chronic Chagas' disease. *Chemotherapy* 55(4):228–233
  64. Chen W, Zeng H, Tan H (2000) Cloning, sequencing, and function of sanF: A gene involved in nikkomycin biosynthesis of *Streptomyces ansochromogenes*. *Curr Microbiol* 41(5): 312–316
  65. Stauffer CS et al (2007) Total synthesis and antifungal activity of a carbohydrate ring-expanded pyranosyl nucleoside analogue of nikkomycin B. *J Org Chem* 72(26):9991–9997
  66. Yamaguchi H et al (1986) Capuramycin, a new nucleoside antibiotic. Taxonomy, fermentation, isolation and characterization. *J Antibiot (Tokyo)* 39(8):1047–1053
  67. Hotoda H et al (2003) Synthesis and antimycobacterial activity of capuramycin analogues. Part 2: acylated derivatives of

- capuramycin-related compounds. *Bioorg Med Chem Lett* 13(17):2833–2836
68. Hotoda H et al (2003) Synthesis and antimycobacterial activity of capuramycin analogues. Part 1: substitution of the azepan-2-one moiety of capuramycin. *Bioorg Med Chem Lett* 13(17):2829–2832
  69. Reddy VM, Einck L, Nacy CA (2008) In vitro antimycobacterial activities of capuramycin analogues. *Antimicrob Agents Chemother* 52(2):719–721
  70. Isono F et al (1989) Mureidomycins A-D, novel peptidyl-nucleoside antibiotics with spheroplast forming activity. III. Biological properties. *J Antibiot (Tokyo)* 42(5):674–679
  71. Isono F et al (1989) Mureidomycins A-D, novel peptidyl-nucleoside antibiotics with spheroplast forming activity. II. Structural elucidation. *J Antibiot (Tokyo)* 42(5):667–673
  72. Inukai M et al (1989) Mureidomycins A-D, novel peptidyl-nucleoside antibiotics with spheroplast forming activity I Taxonomy, fermentation, isolation and physico-chemical properties. *J Antibiot (Tokyo)* 42(5):662–666
  73. Hancock RE, Sahl HG (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat Biotechnol* 24(12):1551–1557
  74. Sahl HG (2006) Optimizing antimicrobial host defense peptides. *Chem Biol* 13(10):1015–1017
  75. Peschel A, Sahl HG (2006) The co-evolution of host cationic antimicrobial peptides and microbial resistance. *Nat Rev Microbiol* 4(7):529–536
  76. Willey JM, van der Donk WA (2007) Lantibiotics: peptides of diverse structure and function. *Annu Rev Microbiol* 61:477–501
  77. Ekkelenkamp MB et al (2005) Isolation and structural characterization of epilancin 15X, a novel lantibiotic from a clinical strain of *Staphylococcus epidermidis*. *FEBS Lett* 579(9):1917–1922
  78. Castiglione F et al (2008) Determining the structure and mode of action of microbisporicin, a potent lantibiotic active against multiresistant pathogens. *Chem Biol* 15(1):22–31
  79. Bauer R, Dicks LM (2005) Mode of action of lipid II-targeting lantibiotics. *Int J Food Microbiol* 101(2):201–216
  80. Cao LT et al (2007) Efficacy of nisin in treatment of clinical mastitis in lactating dairy cows. *J Dairy Sci* 90(8):3980–3985
  81. Wu J, Hu S, Cao L (2007) Therapeutic effect of nisin Z on subclinical mastitis in lactating cows. *Antimicrob Agents Chemother* 51(9):3131–3135
  82. Fernandez L et al (2008) The bacteriocin nisin, an effective agent for the treatment of staphylococcal mastitis during lactation. *J Hum Lact* 24(3):311–316
  83. Nguyen KT et al (2006) Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc Natl Acad Sci USA* 103(46):17462–17467
  84. Penn J et al (2006) Heterologous production of daptomycin in *Streptomyces lividans*. *J Ind Microbiol Biotechnol* 33(2):121–128
  85. Baltz RH, Miao V, Wrigley SK (2005) Natural products to drugs: daptomycin and related lipopeptide antibiotics. *Nat Prod Rep* 22(6):717–741
  86. Miao V et al (2005) Daptomycin biosynthesis in *Streptomyces roseosporus*: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology* 151(Pt 5): 1507–1523
  87. Tally FP, DeBruin MF (2000) Development of daptomycin for Gram-positive infections. *J Antimicrob Chemother* 46(4): 523–526
  88. Henken S et al (2010) Efficacy profiles of daptomycin for treatment of invasive and noninvasive pulmonary infections with *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* 54(2):707–717
  89. Fowler VG Jr et al (2006) Daptomycin versus standard therapy for bacteremia and endocarditis caused by *Staphylococcus aureus*. *N Engl J Med* 355(7):653–665
  90. Walsh C, Freel Meyers CL, Losey HC (2003) Antibiotic glycosyl-transferases: antibiotic maturation and prospects for reprogramming. *J Med Chem* 46(16):3425–3436
  91. Sussmuth RD, Wohlleben W (2004) The biosynthesis of glycopeptide antibiotics—a model for complex, non-ribosomally synthesized, peptidic secondary metabolites. *Appl Microbiol Biotechnol* 63(4):344–350
  92. Donadio S et al (2005) Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis. *Mol Genet Genomics* 274(1):40–50
  93. Levine DP (2006) Vancomycin: a history. *Clin Infect Dis* 42(Suppl 1):S5–S12
  94. Gonzalez C et al (1999) Bacteremic pneumonia due to *Staphylococcus aureus*: A comparison of disease caused by methicillin-resistant and methicillin-susceptible organisms. *Clin Infect Dis* 29(5):1171–1177
  95. Scheinfeld N (2007) A comparison of available and investigational antibiotics for complicated skin infections and treatment-resistant *Staphylococcus aureus* and enterococcus. *J Drugs Dermatol* 6(1):97–103
  96. Maki H, Miura K, Yamano Y (2001) Katanosin B and plusbacin A(3), inhibitors of peptidoglycan synthesis in methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother* 45(6):1823–1827
  97. Fulco P, Wenzel RP (2006) Ramoplanin: a topical lipoglycopeptide antibacterial agent. *Expert Rev Anti Infect Ther* 4(6):939–945
  98. Weissman KJ, Leadlay PF (2005) Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol* 3(12): 925–936
  99. Menninger JR (1995) Mechanism of inhibition of protein synthesis by macrolide and lincosamide antibiotics. *J Basic Clin Physiol Pharmacol* 6(3–4):229–250
  100. McDaniel R et al (1995) Rational design of aromatic polyketide natural products by recombinant assembly of enzymatic subunits. *Nature* 375(6532):549–554
  101. Katz L (1997) Manipulation of modular polyketide synthases. *Chem Rev* 97(7):2557–2576
  102. Ruan X et al (1997) Acyltransferase domain substitutions in erythromycin polyketide synthase yield novel erythromycin derivatives. *J Bacteriol* 179(20):6416–6425



103. Floss HG, Yu TW (2005) Rifamycin-mode of action, resistance, and biosynthesis. *Chem Rev* 105(2):621–632
104. Kim W et al (2009) Rational biosynthetic engineering for optimization of geldanamycin analogues. *Chembiochem* 10(7):1243–1251
105. Bedin M et al (2004) Geldanamycin, an inhibitor of the chaperone activity of HSP90, induces MAPK-independent cell cycle arrest. *Int J Cancer* 109(5):643–652
106. Neckers L, Schulte TW, Mimnaugh E (1999) Geldanamycin as a potential anti-cancer agent: its molecular target and biochemical activity. *Invest New Drugs* 17(4):361–373
107. Delcour AH (2009) Outer membrane permeability and antibiotic resistance. *Biochim Biophys Acta* 1794(5):808–816
108. Mollmann U et al (2009) Siderophores as drug delivery agents: application of the “Trojan Horse” strategy. *Biometals* 22(4):615–624
109. Braun V et al (2009) Sideromycins: tools and antibiotics. *Biometals* 22(1):3–13
110. Ballouche M, Cornelis P, Baysse C (2009) Iron metabolism: a promising target for antibacterial strategies. *Recent Pat Antiinfect Drug Discov* 4(3):190–205
111. Wenciewicz TA et al (2009) Is drug release necessary for antimicrobial activity of siderophore-drug conjugates? Syntheses and biological studies of the naturally occurring salmycin “Trojan Horse” antibiotics and synthetic desferridanoxamine-antibiotic conjugates. *Biometals* 22(4):633–648
112. Borisova SA et al (2010) Biosynthesis of rhizotocins, antifungal phosphonate oligopeptides produced by *Bacillus subtilis* ATCC6633. *Chem Biol* 17(1):28–37
113. Vondenhoff GH et al (2011) Characterization of peptide chain length and constituency requirements for YejABEF-mediated uptake of Microcin C analogues. *J Bacteriol* 193(14):3618–3623

## Aquaculture and Renewable Energy Systems, Integration of

BELA H. BUCK<sup>1,2,3</sup>, GESCHE KRAUSE<sup>4</sup>

<sup>1</sup>Biological Sciences – Marine Aquaculture, Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven, Germany

<sup>2</sup>Marine Aquaculture for Sustainable Fisheries, Institute for Marine Resources (IMARE), Bremerhaven, Germany

<sup>3</sup>Maritime Technologies – Applied Marine Biology, University for Applied Sciences, Bremerhaven, Germany

<sup>4</sup>Leibniz Center for Tropical Marine Ecology (ZMT), Bremen, Germany

## Article Outline

Glossary

Definition of the Subject

Introduction

Status Quo of Offshore Aquaculture Research Activities in Wind Farms

Future Directions

Bibliography

## Glossary

**Aquaculture** Following the definition of the FAO [1, 2], *aquaculture* is the farming of aquatic organisms, including fish, molluscs, crustaceans, and aquatic plants with some sort of intervention in the rearing process to enhance production, such as regular stocking, feeding, and protection from predators. Specifically, *marine aquaculture*, also called *mariculture*, concentrates on aquatic organisms cultivated in brackish or marine environments.

## Integrated Coastal Zone Management (ICZM)

A process for the management of the coast using an integrated approach, regarding all aspects of the coastal zone, including geographical and political boundaries, in an attempt to achieve sustainability. The EU Commission [3] defines ICZM as a dynamic, multidisciplinary, and iterative process to promote sustainable management of coastal zones. It covers the full cycle of information collection, planning (in its broadest sense), decision making, management, and monitoring of implementation. ICZM uses the informed participation and cooperation of all stakeholders to assess the societal goals in a given coastal area, and to take actions towards meeting these objectives. ICZM seeks, over the long term, to balance environmental, economic, social, cultural, and recreational objectives, all within the limits set by the natural dynamics.

**Mariculture** See “aquaculture”.

**Offshore aquaculture** A culture operation in a frequently hostile open ocean environment exposed to all kinds of sea states as well as being placed far off the coast.

**Offshore co-management** A dynamic partnership using the capacities and interests of different stakeholder groups for managing cross-sectoral activities

in cooperation with governmental authorities in the open sea.

**Offshore wind farms** A group of wind turbines in the same confined area used for production of electric power in the open ocean. Moving off the coast to the offshore, wind turbines are less obtrusive than turbines on land, as their apparent size and noise is mitigated by distance. Since water has less surface roughness than land (especially in deeper waters), the average wind speed is usually considerably higher over the open water. Therefore, the capacity factors are considerably higher than for onshore and nearshore locations [4].

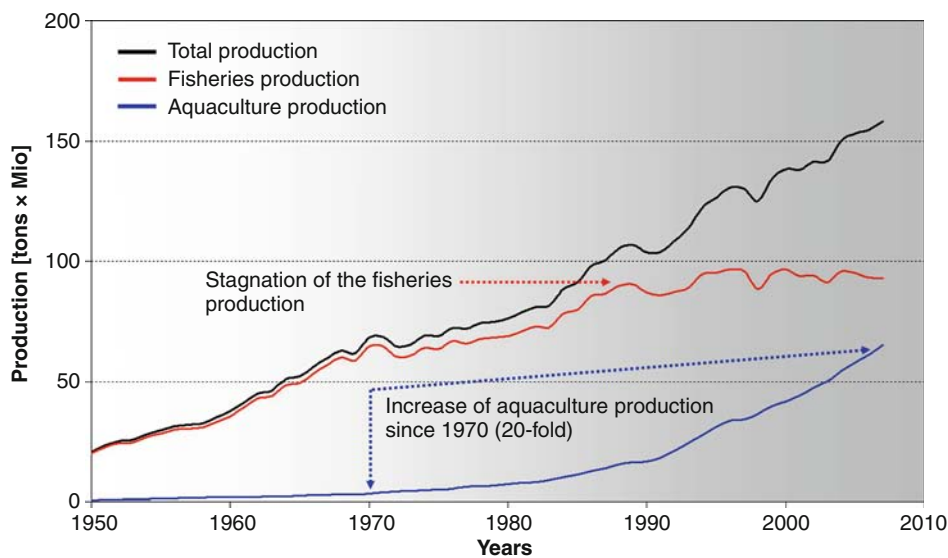
**Open ocean aquaculture** See “Offshore aquaculture”.

### Definition of the Subject

“Fisheries have rarely been sustainable.” This statement by Pauly et al. [5] was based on the recognition that this lack of sustainability was induced by a serial depletion of wild stocks worldwide. Causative for this trend is due to the improved fishing technology, geographical expansion, and exploitation of previously spurned species lower in the food web. In exchange, aquaculture

was often either regarded to bridge the gap between supply and demand or, in contrast, even to exacerbate this scenario.

Since the 1970s, aquaculture production has grown quite rapidly and is by now one of the fastest growing aquatic food production sectors in the world [6]. Besides the rapid development of this sector, the wide-ranging decline in fisheries yields has been enhanced by an increase in public demand for aquatic products. With an annual share of more than 15% of total animal protein supplies, the production of captured fisheries and aquaculture plays a significant role in the global food security [6]. In 2007, approximately 160 million tons of aquatic organisms were produced worldwide (Fig. 1). From that, the share of global aquaculture production amounts to almost 47%, totaling about 60 million tons annually of aquatic organisms [7, 10]. A wide range of aquatic species is raised in various systems, onshore as well as in the ocean. According to the FAO [6], approximately 300 different species, ranging from fish to shellfish, crustaceans and algae are produced in aquaculture systems. Most of these aquaculture enterprises are concentrated in well-protected and therefore favorable inshore water areas [11].



**Aquaculture and Renewable Energy Systems, Integration of. Figure 1**

Global production of aquatic organisms originating from fisheries and aquaculture within the last 55 years (Data source [7], modified after [8, 9])

Even probably though over-reporting its aquaculture production [12], the People's Republic of China has contributed approximately 70% to the world's aquaculture production in 2004. It is nevertheless debatable, whether this production can compensate for the global deficiency in aquatic food. In addition, the intensive traditional aquaculture of carnivorous species does not automatically relieve pressure on ocean fisheries [13]. Salmon farming, e.g., requires large inputs of wild fish as fish oil and fish meal for the production of fish feed for aquaculture. Hence, the farming of non-carnivorous species that is not dependent on fishmeal-based feeds is considered a sustainable way of producing food. However, the global increase in production originates from herbivorous species. Further, the balance between carnivorous and non-carnivorous species in aquaculture production is heavily skewed towards herbivorous species [14].

On top of this issue, an increasing limitation of favorable coastal sites for the development of modern aquaculture is evident in various countries, such as Germany, the Netherlands, Belgium, as well as others [15]. This spatial limitation is mainly caused by the lack of protected nearshore areas and by the fact that regulatory frameworks that assign specific areas for aquaculture operations are diverse and still emerging.

Further, the utilizations of coastal marine waters are manifold and quite competitive, such as shipping (trade or private), recreational activities, extraction or disposal of gravel, marine missions, fisheries, mariculture, offshore wind farms, cable and pipelines, establishment of nature reserves, and other marine and coastal protected areas. In addition, overlapping use of coastal habitats adds to the increasing pollution of coastal waters in various situations and gives rise to spatial conflicts, thus leaving little room for the expansion of modern coastal aquaculture systems.

This situation in most industrialized countries is often in contrast to the production progress in developing countries. Here, the installation of aquaculture systems benefits from the often weak enforcement of integrated coastal management schemes, which regulate equal access to the coastal resources [16, 17]. Thus, the rise of aquaculture production has specifically taken place in developing countries, especially in Asia, which holds approximately 91.4% of the global production share [10, 18, 19]. In contrast, the number of competing

users within offshore regions is relatively low, thus favoring the offshore environment for further commercial development, such as offshore wind farming and open ocean aquaculture. So far, spatial regulations offshore are scarce and clean water can be expected [20]. Thus, there is an enormous economic potential for *extensive marine aquaculture* in offshore areas.

## Introduction

Aquaculture has been increasing dramatically in most parts of the world and now accounts for more than 47% of the total global seafood supply [7]. Many people generally assess aquaculture positively as a potential alternative to global fishery resources, which are globally under stress as a result of overfishing. However, it also raises concerns over pollution, disease transmission, and other socio-economic impacts. Almost all efforts to develop marine aquaculture have focused on state jurisdictional waters of the coastal sea, which are generally situated within 3 nautical miles off the shore [21, 22]. With the convergence of environmental and aesthetic concerns, aquaculture, which is already competing for space with other more established and accepted uses, is having an increasingly difficult time expanding in nearshore waters [23]. Therefore, alternative approaches are needed in order to allow the expansion of the marine aquaculture sector to make a meaningful and sustainable contribution to the world's seafood supply.

The political recognition – on a national as well as on EU level – that the implementation of integrated coastal zone management (ICZM) is still fragmentary, acted as incentive to investigate in more detail how this could be overcome [e.g., 24]. This lack was recognized and led to the operation of a EU-Demonstration Programme on Integrated Coastal Zone Management from 1996 to 1999. This Programme was designed around a series of 35 demonstration projects and six thematic studies. In 2002, based on the experiences and outputs of the Demonstration Programme, the EU-Commission adopted a recommendation concerning the implementation of Integrated Coastal Zone Management in Europe (Recommendation of the European Parliament and of the Council, 2002/413/EC). In Germany, this generated a call of the Federal Ministry of Education and Research to the various federal states to

develop projects that address ICZM on a regional level. In 2004, the program *Coastal Futures* [25], which tied up various administrative and scientific bodies and the public along the west coast of the State of Schleswig-Holstein, was granted funding. This program focused primarily on two issues: [1] to develop the future of the coast as a living, working, and recreational space for the local population, and [2] to consider the potential contribution of coastal resources to the sustainable development on the national and EU/global level, i.e., by providing regenerative energy by wind power. In order to sustain sufficient open space for future development, the idea of combining offshore wind power generation with other uses, such as aquaculture operations, emerged [26]. Marine aquaculture is a growing enterprise in Germany as well as in the whole of Europe, strongly motivated by the decline of fisheries production and the search for alternative income options for rural peripheral coastal regions.

In order to stimulate multifunctional use of marine space, it was decided to develop a project on a showcase basis, which deals not only with different scientific fields but also with private–public partnerships and the relevant institutional bodies. In the following, an overview on the current state of research undertaken within this focus is provided. Offshore wind farms will hereby act as a case example for renewable energy systems in the open ocean.

### Offshore Aquaculture – A New Addition to Marine Resource Use

Farming in the open ocean has been identified as one potential option for increasing seafood production and has been a focus of international attention for more than a decade. *Offshore aquaculture* or *open ocean aquaculture* are operations in a marine environment fully exposed to all kinds of oceanographic conditions [27] as well as located at least 8 nautical miles off the coast [15] to avoid the many stakeholder conflicts in nearer coastal areas [28]. The procedures and applied techniques for the cultivation of organisms mainly depend on the species; their life cycle determines the phases of cultivation and the location for the grow-out, where market size will be reached. First trials of cultivation were based on *extensive marine aquaculture*, which – in contrast to intensive aquaculture – is a line of

production with little impact on the marine environment. These aquaculture operations are characterized by (1) a low degree of control (i.e., environmental control, nutrition, predators, competitors, and disease agents), (2) low initial costs, (3) low level technology, (4) low-production efficiency, and (5) high dependence on local climate and water quality (natural water bodies, such as bays, ponds, embayments) [29]. Mostly, they are regarded as a sustainable line of production.

Moving to the open ocean has been considered as a means for moving away from negative environmental impacts and negative public perception issues in the coastal zone. Favorable features for the transfer to open ocean waters include ample space for expansion and thus reduced conflicts with other user groups, lower exposure to human sources of pollution, the potential to reduce some of the negative environmental impacts of coastal fish farming, and optimal environmental conditions for various marine species through the larger carrying and assimilative capacities. However, this move should not be seen as an “out of sight, out of mind” attitude, as open ocean development will also come under scrutiny by the institutional bodies as well as by a more and more educated public. It is expected that, because of economies of scale, the open ocean farms of tomorrow will be larger than the present nearshore farms. Therefore, higher levels of waste can be generated. Even if greater residual effects occur, deeper waters and lower nutrient baselines are expected to reduce impacts from open ocean operations through wider dispersion plumes of nutrients, as compared to similarly sized nearshore operations. However, there will be a point when open ocean ecosystems will eventually reach their assimilative carrying capacities [30].

### Offshore Wind Farms as a Case Example for Renewable Energy Systems

Wind energy continues to be the world’s most dynamically growing energy source [31]. Drawing on the example of Germany, the first initiative toward an economy based on renewable energy resources was set by the governmental decision in the year 2000 to gradually reduce the use of nuclear energy and to respond to the gradually diminishing fossil- and nuclear-energy reserves. Simultaneously, the output of CO<sub>2</sub> to the

atmosphere would be reduced in accordance with the Kyoto protocol as well as the dependence on conventional fossil-energy resources is lowered.

As high and reasonably steady wind speeds are characteristic in Northern offshore areas, these areas are prime candidates for renewable energy production by wind-energy farms. For instance in the North Sea, a major political incentive exists currently to install large offshore wind farms [32, 33]. Thus, the emerging branch of offshore wind farms appears as a new stakeholder on the list of users [34, 35].

So far, this development has been successful to such an extent that around 7.2% of the total energy consumption in Germany is covered by this technology. At the end of 2007, Germany had an installed capacity of 22,247 MW, generated by 19,460 mainly land-based operating wind turbines [36]. Within Europe, as the leading market for wind energy with over 57 GW, Germany thus accounted for 39% in terms of the total installed capacity and still remains the world's leader. However, with the North American market currently experiencing a strong growth, it is expected that the US market will soon overtake Germany [37].

At present, 60 project applications for wind farms in the Exclusive Economic Zone (EEZ) of the German North Sea and in the Baltic Sea are in the planning process stage with the total number of wind turbines per farm ranging between 80 and 500 [26] (Fig. 2). In November 2001, the Federal Maritime and Hydrographic Agency (BSH) granted the first approval for the installation of a pilot offshore wind farm. Since then, a total of 23 wind farm development projects have been approved in German waters, most of them planned seaward of the 12 nautical miles zone [38]. Currently, a larger test farm of about 12 wind turbines (5 MW class) at the "Borkum West" site are in operation (Fig. 3) [39]. Experience gained in this project should give developers practical knowledge in the construction and operation of offshore wind farms at depths (down to 50 m) and at distances from the shore (up to 50 nautical miles and more) that are beyond comparison to those anywhere in the world [31, 33].

In contrast to neighboring European states, the prospect of moving wind energy developments offshore stagnated in Germany for years mainly due to a very complex licensing procedure and the high environmental constraints [33, 40]. A further obstacle roots

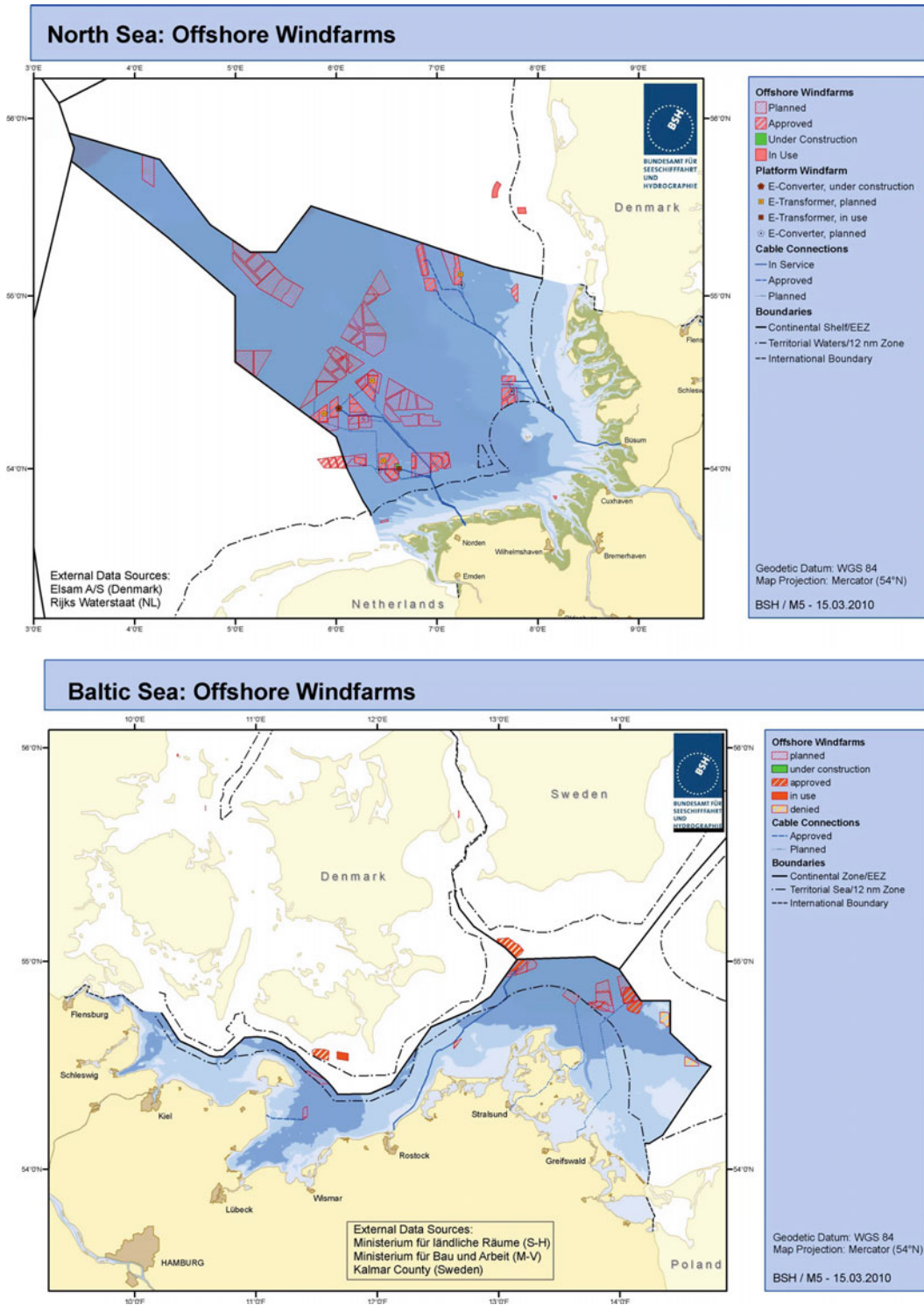
in the spatial competition of offshore wind farms with other utilization of the marine waters in the German Bight [41, 42]. However, despite the number of competing users within offshore regions being lower compared to coastal areas [43], the quest for spatial efficiency remains to be a key incentive also for offshore developments in the future.

### Moving Offshore: The Multiple-Use Concept

The plans for the massive expansion of wind farms in offshore areas of the North Sea triggered the idea of a combination of wind turbines with installations for extensive shellfish and macroalgae aquaculture [15, 26]. Offshore wind farms provide an appropriately sized area free of shipping traffic as most offshore wind farms are designed as restricted-access areas due to hazard mitigation concerns. Concurrently, the infrastructure for regular service support is readily available, and hence such sites provide an ideal opportunity for devising and implementing a multiple-use concept [42, 44]. However, in contrast to coastal inshore areas where beaches and their adjacent near-shore zones act as buffers to absorb wave energy, offshore regions are high-energy environments, fully exposed to waves, weather, and currents. Numerous studies have demonstrated that waves can reach remarkable heights (up to 10 m) in the offshore areas of the North Sea [e.g., 45, 46]. In this context, the solid foundation structure of wind turbines provides support for anchoring cultivation devices that can withstand the harsh oceanic conditions [47]. Furthermore, offshore structures are well known for their artificial reef function, thus supporting biodiversity in the ecosystem. The offshore water quality, which is a major issue in all kinds of aquaculture operations, is regarded to be excellent in comparison to inshore areas [48, 49]. Finally, the multifunctional use of offshore areas reduces conflicts between stakeholders if activities are concentrated and conjointly managed within so-called multiple-use marine areas. This, in turn, increases the amount of open ocean territory free of utilization by man. All of the above issues are considered as key incentives to move offshore with aquaculture operations.

In view of the many interests for the offshore move, different suggestions for technical structures for open





**Aquaculture and Renewable Energy Systems, Integration of. Figure 2**

Maps indicating all application sites for wind farm projects in Germany. At the top, the North Sea, below the Baltic Sea areas (Modified after [38])



**Aquaculture and Renewable Energy Systems, Integration of. Figure 3**

Offshore wind farm Alpha Ventus. (a) Shows the transfer of the windmill tripods to the harbor of Wilhelmshaven and (b) displays the setup of an offshore windmill (REpower MI 068 [39])

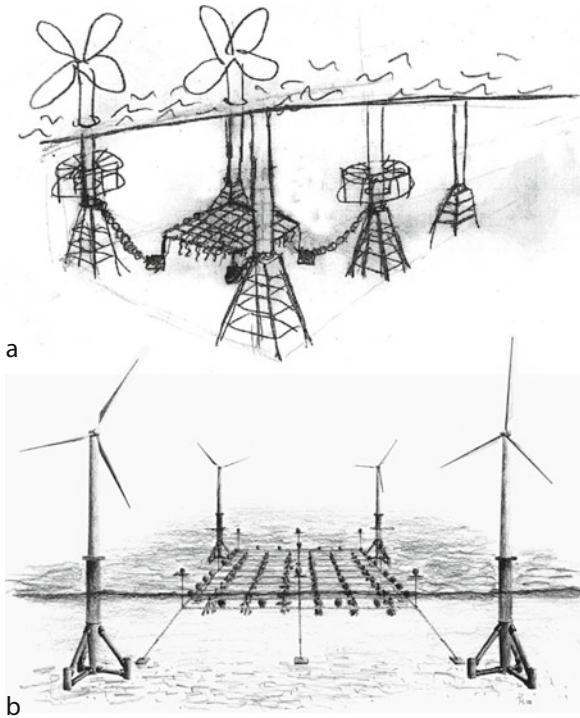
ocean aquaculture were proposed (see proceedings of various OOA-Conferences [e.g. 50, 51]), which could cope with the harsh environmental conditions that place an enormous stress on the employed materials. It would be advantageous for the global offshore aquaculture development to plan for a combination of uses. While windmills use the wind above the surface to produce energy, their fixed pylons, commonly concrete fundaments (gravity foundation), metal jackets, tripods, or triples offer a possibility to connect systems used in aquaculture (Fig. 4). The combination of the respective two industries has to cope with the forces generated by the high-energy environment.

Since 2000, when the co-use of wind farms for off-bottom offshore cultivation [26] in the German Bight was proposed, several studies have been conducted to elucidate the potential as well as constraints of this offshore alternative for extensive aquaculture. Two pioneer studies, the project *Roter Sand* and *Offshore Aquaculture* were conducted between 2002 and 2004 by the Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany. These two projects followed

a complex approach to obtain data about suitable indigenous candidates for the offshore cultivation [15], the technical requirements of longline systems for the cultivation of mussels or oysters [8, 9] and algal cultivation systems [53]. Insights into the feasibility of offshore seed and mussel production concerning larval, nutrient, and phytoplankton concentrations [8, 9, 54] were provided, and the existing legislation and regulations concerning marine aquaculture in Germany were listed [21]. In addition, all stakeholders potentially involved in a multifunctional use of offshore wind farms for aquaculture were identified [42]. This successful multifaceted approach helped to disperse many concerns and doubts on the offshore idea and initiated a sequence of and relations between various following projects, which are displayed in Fig. 5.

#### Candidates and Techniques for the Multi-Use Concept

In general, the cultivation process should consider only indigenous species for marine aquaculture operations to



**Aquaculture and Renewable Energy Systems, Integration of. Figure 4**

Potential multifunctional use of fixed underwater structures of wind turbines for the operation of aquaculture facilities: 12 years ago and today (2010).

(a) First drawing ever for the multi-use concept, including alternative solutions of oyster cages and mussel collectors attached to longlines in the inner section of the wind farm or offshore-rings (collar systems) attached directly to the pylon. The latter system can be submersed in case of wind-turbine maintenance. (b) Presents a design of a single mussel plot within a group of four wind turbines (not to scale) (Modified after [52], Buck personal drawing)

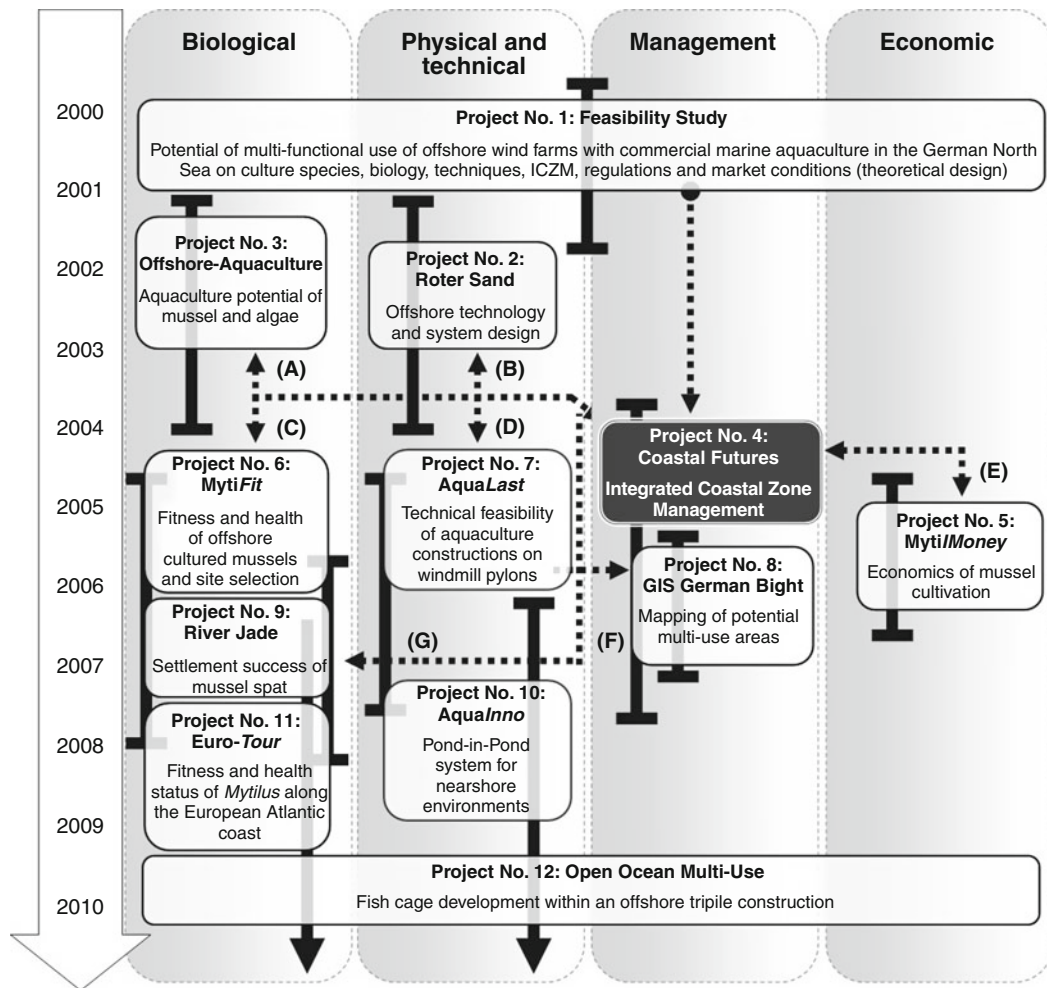
avoid the disruption with the local marine flora and fauna. This limits the economic opportunities of marine aquaculture enterprises since in certain sites only a few indigenous candidates are regarded as high-value species. Following a feasibility study by Buck [26], in Germany only culture species with modest service needs can be considered as favorable candidates for offshore aquaculture. In the offshore test trials in Germany, most suitable candidates suggested and tested were the sugar kelp (*Laminaria saccharina*), oarweed (*L. digitata*), dulse (*Palmaria palmata*), the blue mussel (*Mytilus*

*edulis*), and two oyster species, the Pacific oyster (*Crassostrea gigas*) and the European flat oyster (*Ostrea edulis*). Mussels and seaweeds, for example, are cultured mainly in extensive systems throughout the world [8, 56, 57]; the latter occurs for historical and traditional reasons mostly in Asian countries.

According to Tseng [58], the cultivation procedure of brown seaweeds can be divided into two separate steps: In step (1), the seedling phase, spores are artificially released from mature sporophytes and seeded on a given substrate (ropes wrapped around plastic frames), where germination of gametophytes, the sexual maturation of male and female gametophytes, and finally, the development of zygotes into juvenile sporophytes takes place. In step (2), the grow-out phase, culture ropes with juvenile sporophytes are transferred to the open sea. In the grow-out phase, the macroalgal sporophytes grow on ropes for one season to a frond length of approximately 2 m.

When natural reproduction of mussels occurs, gametes are released into the water column where fertilization takes place [59]. The larvae undergo all trochophore and veliger stages when settling on a given substrate to start metamorphosis. According to Pulfrich [60] and Walter and Liebezeit [61], this process normally takes place at spring time (larval peak in May) in the German Bight. The cultivation of blue mussels can be divided into two steps: in step (1) the naturally occurring spat collection is achieved by deploying artificial substrates [62]. Usually, spat collectors are made out of unraveled polypropylene lines or sisal ropes, to offer the mussel's post larvae substrate for settlement [56]. After several months (step 2), collectors are retrieved and mussels thinned out and reseeded on ropes to provide space to improve growth and allow fattening [63, 64].

To operate culture phase (2) of both species, macroalgae and bivalves, an appropriate system design, such as suspended longlines or floating ring-structures, have to be deployed and securely moored in order to resist the stress forces of incoming waves and tidal currents, as well as swell. In addition, it was necessary to assess what kind of technical structure supports best the growth of the organisms (e.g., prevention from loss or mortality) while also assessing whether such systems provide reasonable production returns. Finally, potential combinations with offshore wind turbines had to be assessed.



**Aquaculture and Renewable Energy Systems, Integration of. Figure 5**

Chronological order of conducted and ongoing research projects dealing with the combination of offshore wind farming and open ocean aquaculture. Project No. 1, the feasibility study, constituted the basis for all subsequent research. The *Coastal Futures Project* acts as a key node project to which the other projects either have contributed or by which they have been stimulated because of its transdisciplinary approach. It is visible that: (a) calls the wind farm developers' attention to offshore aquaculture; (b) and (c) include authorities and fishermen into the planning process for site-selection criteria of appropriate aquaculture sites; (d) involves offshore engineers and wind farm developers/operators into the technical part of an offshore aquaculture enterprise; (e) introduces (mussel) fishermen to the co-management idea and appraises the economics of mussel cultivation; (f) supplies authorities with maps and tools to limit regional stakeholder conflicts, (g) establishing an inshore reference station to support the data collected offshore, and (h) testing the first fish cage mounted within a tripile construction (Modified after [55])

However, currently even candidates requiring a semi-intensive as well as intensive cultivation process are in the testing phase. Salmon (*Salmo salar*), seabass (*Dicentrarchus labrax*), seabream (*Sparus aurata*), or some flatfish species are discussed for aquaculture in

fish cages below windmill platforms at different offshore sites worldwide. Fish will firstly be reared in land-based facilities and will then be transferred as fingerlings to the offshore site and released into the submersible fish cages. After reaching market size, the



fish will be harvested and removed to the land and will undergo normal processing procedures.

Relocating cultivation systems offshore into high-energy environments requires the development of suitable culture techniques able to withstand the harsh conditions and minimize risk of economic loss [65].

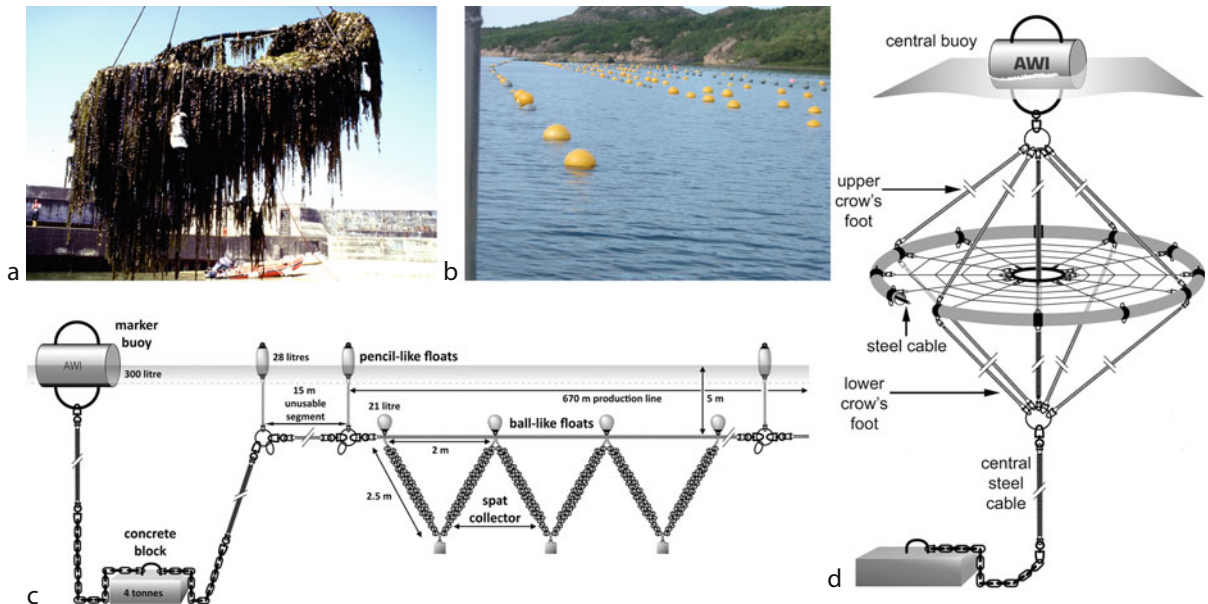
Several techniques exist to cultivate mussels and seaweed either in co-culture or in single culture. Basically, both organisms are cultured in a suspended manner in the water column, floating or submerged. The use of rafts, longlines, and ring methods dominate. The latter two were the main cultivation techniques used in test trials offshore wind farm areas [8, 53, 56] (Fig. 6).

Major difficulties in the development of suitable techniques for open ocean aquaculture are – as mentioned above – the harsh environmental conditions which place an enormous stress on materials. Depending on the acting hydrodynamic forces, different technical setups can be distinguished. One of the interesting possible linkages of aquaculture is the combination with offshore wind farms as these would

provide stable fixing structures for the cultivation systems. This is especially relevant from an economic point of view as so far the costly infrastructure for offshore aquaculture systems is one of the major drawbacks in the development.

### Status Quo of Offshore Aquaculture Research Activities in Wind Farms

Only a few scientific studies dealing with the prospects of offshore aquaculture were available before 2000, and little was known about the biotechnological requirements, economic potential, or the socio-economic influence on the general feasibility of offshore aquaculture. Very few long-term experiments under harsh hydrodynamic conditions exist, e.g., Langan and Horton for offshore mussel cultivation [66]; Neushul and Harger [67]; Neushul et al. [68] for offshore seaweed cultivation. However, data on system and species performance are urgently needed to derive methodologies for the assessment of its environmental and economic viability. Therefore,



**Aquaculture and Renewable Energy Systems, Integration of. Figure 6**

Aquaculture constructions suitable for the cultivation in high-energy environments. (a) Offshore ring design for the cultivation of macroalgae (here: harvesting after grow-out in the harbor of Helgoland), (b) example of a nearshore, submerged longline design for mussels and oysters, (c) schematic drawing of a submerged longline suitable for exposed sites, and (d) a technical illustration of the ring design and its mooring system (Modified after [8, 9, 53])



the assessment of the potentials and constraints for sustainable aquaculture development in all marine habitats requires input from various scientific disciplines in order to direct this development towards a successful aquaculture undertaking. In particular, this holds true for offshore aquaculture, where little practical experience is available to date, although research in this area is evolving rapidly (e.g., Buck [15], Turner [69]; Pérez et al. [70]; Bridger and Costa-Pearce [51]; Dalton [71]; Naylor and Burke [72]).

The offshore wind farm and aquaculture investigations initiated an integrated assessment of theoretical and practical challenges of aquaculture operations in the North Sea. Several studies were carried out, all of which contributed to specific aspects of such a combined utilization of offshore space. These were:

(a) *Biological studies*, in which the focus was placed on cultivation and subsequent performance characteristics of indigenous bivalve, seaweed, and fish species exposed to extensive offshore aquaculture farming conditions. Further, the health status and infestation rates with parasites, bacteria, and viruses of candidates were determined to gain reliable predictions on where the highest growth rates and best product quality for consumers can be achieved. In nearshore intertidal areas, mussels and oysters are particularly exposed to high concentrations of pollutants, pesticides, near surface agents, estuarine run-offs, etc. that can pose a threat to consumer health. Buck [8, 9, 15] reported high growth rates for mussels cultivated in the German Bight. The scope of growth, i.e., the energy available for growth, is usually directly and positively correlated to a good overall health condition of the respective organism [73]. But organisms with high growth rates and a healthy appearance are no guarantee of a healthy food for human consumers. For instance, in coastal waters, eutrophicated by urban sewage, mussels show good growth performance. The microbial status of these mussels, however, mostly excludes them from consumption since they might carry various human pathogens. Even in developed countries with strict legislation for the treatment of wastewater, mussels can function as carriers of vector diseases. Whether this is also true for offshore cultivated mussels, where the environment is

cleaner due to dilution of contaminants, remains open. Data for offshore-produced mussels, generated according to the analysis protocols of controlling authorities, are not readily available for all cultivation sites. However, new regulations are in the implementation process in all of the EU states and will fulfill the prerequisites for an official sampling design and assessment (i.e., sanitary survey).

To evaluate the significance and comparability of the employed parameters in a broader geographical context, the area of investigation was extended along the Atlantic coast from southern Portugal to northern Denmark. Further on, the closely related Mediterranean mussel *Mytilus galloprovincialis* was included in the analysis to test the effectiveness of all the parameters in different species.

Investigations on fish species for submerged cage-systems included aspects on growth, welfare, stress in exposed environments, and health.

(b) *Physical and technical studies* investigated the effects of the prevailing hydrodynamics on candidates and culture constructions at specific offshore sites. At the same time, the necessary technical requirements for farming structures in high-energy environments and their possible combination with offshore wind farms were assessed. New system designs for offshore farming were developed and prototypes (e.g., offshore ring, offshore collector) were tested. Technical details about the microstructure of artificial substrates were addressed to increase production per meter longline under offshore conditions. In addition to offshore seaweed and mussel cultivation, new technologies for submerged fish cages were investigated.

(c) *Management and institutional studies* focussed on the analysis of potential management approaches to implement a multi-use concept of offshore areas. Hereby, the various stakeholders and their respective views and knowledge systems were integrated. Against the background of the social and institutional dimensions, particular emphasis was given to the interrelationship between scientific findings on the one hand and effective implementation on the other. Key aspects included the social acceptance of combined use, as well as the possible management strategies that would govern it. This endorsed the examination of the prevailing case laws and regulative and management framework

conditions, as well as a suggestion of decisive offshore co-management strategies to support such activities. In this process, the continuous inclusion of the stakeholders in a participatory manner was a prerequisite. To address the respective technical, economic, social, and political challenges of mariculture and offshore wind farms, specific co-management strategies were elaborated that are either more results-oriented (e.g., for integrating technical knowledge of the two sectors) or more process-oriented (e.g., for establishing new linkages between different groups). Thus, in cooperation with governmental authorities, co-management in the offshore makes use of the capacities and interests of the respective stakeholder groups and employs these in managing cross-sectoral activities.

- (d) *Economic studies* conducted an economic evaluation of such multi-use concepts in offshore locations that take into consideration market conditions as well as investment and operating costs.

All the above listed conceptual approaches relied on results of a theoretical feasibility study (Fig. 5) [26], which was carried out prior to practical research in the field. All of the results contribute to the *Coastal Futures* Program and support the quest to find innovative new approaches for sustainable use and alternative livelihoods of coastal populations.

### Overview of Biological and Technical Investigations

Over the last decades, substantial insights have been gained on the terms and conditions active in the offshore environment. However, these data are only partly useful for the selection of offshore aquaculture sites because they have been gathered primarily for other user needs and thus lack the essential specificity to address the biological and cultivable potential of these sites. Prior to a multifunctional development comprising mariculture activities, it is therefore necessary to determine the appropriate biological, technological, and management requirements, as well as the performance characteristics that would allow the employment of favorable and cost-effective methodologies. To meet this end, special focus was placed on the combination of extensive offshore shellfish, seaweed, and fish farming at exposed sites within the proposed offshore wind farm boundaries.

Due to the wide spectrum of open questions, the outcomes are quite manifold. In the following, first results according to their contributions towards the main research topics involved are presented.

**Biological Studies** The theoretical *Feasibility Study* [13, 24] was aimed to ascertain the biological, technical, and economic feasibility of an offshore marine aquaculture structure with respect to the cultivation of marine organisms within wind farm sites in the German North Sea. One result was that to date, in terms of commercial marine aquaculture, Germany had little knowledge and background in offshore aquaculture compared to many other coastal countries throughout the world. Nevertheless, a synthesis of a selection of parameters (e.g., geo-physical and biological parameters) allowed the identification of suitable candidates for commercial offshore aquaculture. These candidates include blue mussels (*Mytilus edulis*) and oysters (*Ostrea edulis*, *Crassostrea gigas*), which could be maintained extensively in the offshore region. Moreover, labor requirement for these candidates as well as for seaweeds, such as the sugar kelp (*Laminaria saccharina*) and dulse (*Palmaria palmata*), is supposed to be low.

Further, the biological feasibility of cultivating mussels, oysters, and kelp within offshore wind farm sites was assessed. The growth of these species is excellent in the rather eutrophicated offshore environments of the North Sea, but can differ depending on exposure sites, system designs, installation modes, and season.

For instance, settlement of young mussels on artificial collector substrates decreases with increasing distance from the shore [74]. However, this does not limit the economic potential if the thinning procedure will be omitted, following a “One-Step-Cultivation” concept [15]. In general it was found that mussels are free of parasites at offshore locations due to dilution effects and the interrupted reproduction cycles of some macroparasites [75]. Special focus was placed on the overall health status of mussels cultured under different conditions, and the impact on economic aspects was investigated [76]. Specific aims of the projects were the development of suitable offshore spat collecting techniques, detailed knowledge about parasites (macro and micro), bacteria and virus infestations at different sites, implementation of biodiagnostic techniques for the

health analysis of cultured mussels, and collection of all relevant data (e.g., shell stability and attachment strength of mussels), for the further processing of mussels as a product for human consumption.

Hydrodynamic forces support length increase of seaweed blades when transferring young sporophytes to sea. These algae will adapt to the occurring loads and develop strong holdfasts, preventing detachment of the entire plant [77].

Modified and improved techniques for offshore farming withstand high-energy environments, but will certainly cause higher investment costs. Therefore, site-selecting criteria for a culture area should be clearly identified to assess economic risks. Important for the cultivation success is the water quality. The analysis of the cultured organisms with biodiagnostic tools provides detailed insights into the water conditions the animals live in. By this approach, reliable predictions are possible as to which locations grant highest growth rates and best product quality for consumers. Preliminary results attest offshore areas satisfying settlement success and excellent growth rates [78], and low infestations with macroparasites [79], microparasites, bacteria, and toxins [76]. The results on consumption suitability show that water quality regarding the concentrations of pollutants in offshore areas of the German Bight is quite good. Lysosomal membrane stability is mostly relatively low at all tested nearshore and offshore sites. Interestingly, growth rates of the hanging cultivated mussels are not affected by this low fitness parameter [58].

First results on investigations along the Atlantic Coast show that mussels originating from offshore habitats have a better health status regarding the infestation with macroparasites and microparasites (Buck and Brenner, unpublished data). While macroparasites are still infesting mussels in nearshore areas in the Wadden Sea (the Netherlands, Germany, Denmark), microparasites are absent.

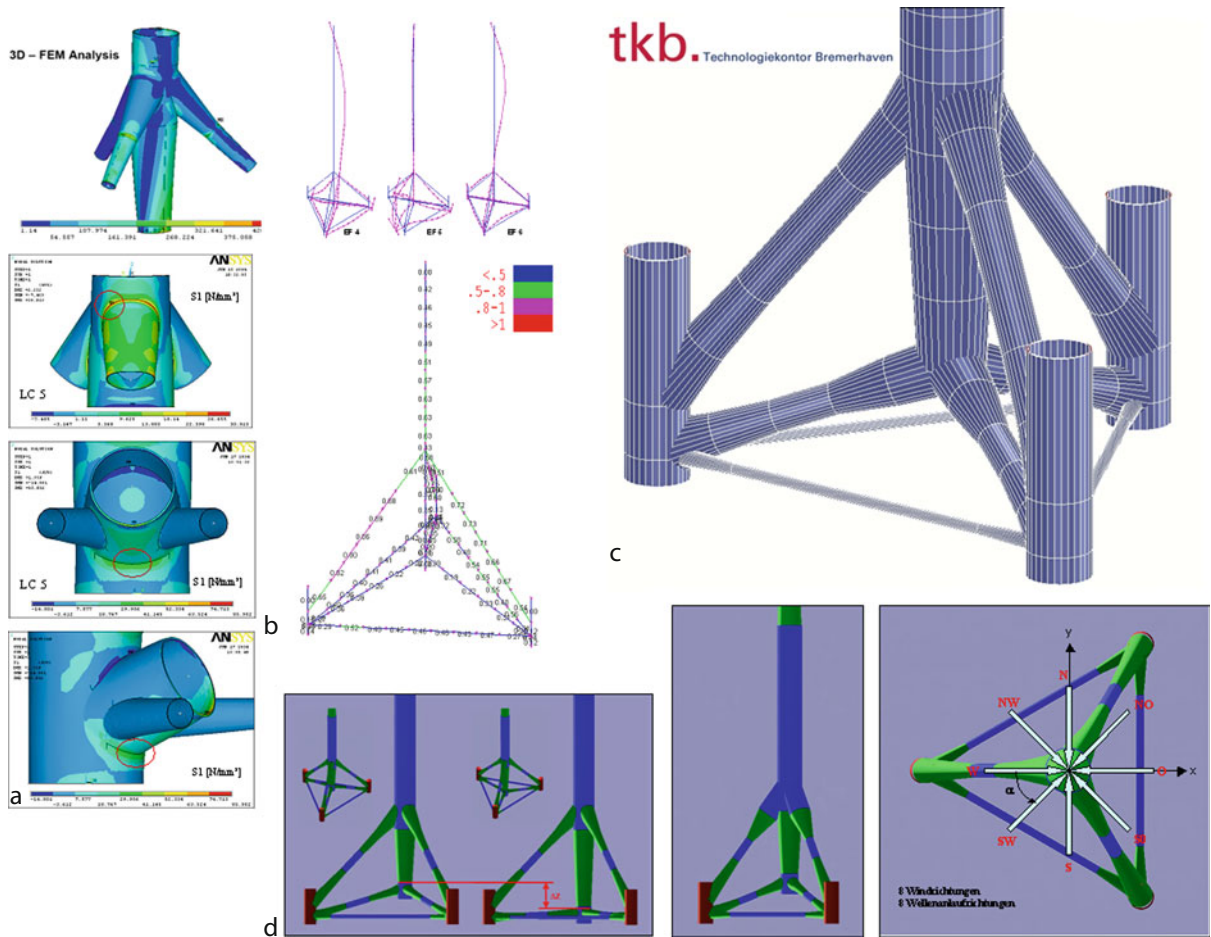
**Physical and Technical Studies** The results above allowed the identification of two offshore aquaculture systems that were best suited for offshore operations from a biological point of view. Depending on the acting hydrodynamic properties, different technical setups are regarded as favorable. The first one is a floating and submergible ring system for the cultivation of seaweed.

It withstands rough weather conditions and allows easy handling [53]. The second system is a submerged longline design for blue-mussel culture [8]. The longline should ideally be installed 5 m below the water surface and should be connected to foundations of offshore windmills (Fig. 7) [47]. For the longline, polypropylene proved to be an appropriate material. The system design is made of various connected segments allowing an easy harvest and replacement of all parts of the construction. However, more technical engineering research is required to find the most cost-effective mode of construction and the best choice of materials (e.g., little corrosion, longevity in spite of mechanical stress) so that easy handling can be guaranteed under relatively harsh weather conditions (cf. construction, deployment, retrieval, service, repairs).

The experimental design also allowed work on such issues as the efficiency of the collecting devices themselves. Healthy mussels will reach market size in offshore conditions only if they are firmly attached to their artificial substrate. As mussels growing on suspended substrates need about 15 months [8, 9] on average to reach market size, they must survive one winter and withstand storm events producing wave heights up to several meters. Continuing investigations on the health and quality of market-sized mussels would be moot if mussels failed to stay attached to substrate gear.

To date, most available substrates are designed and deployed for nearshore use under calm water condition. However, it was found that improvement for construction of new collectors that are feasible for offshore cultivation is in mandate. Research showed that new substrates should have felt-like structures around the core of a collector for larval attraction and long appendices in high density to interweave the mussel conglomerates with the substrate [80]. Future investigations should focus further on the fabrication and testing of a prototype of this collector, concerning the results of this study. Besides providing optimal larval attraction and attachment for juvenile mussels even under winter conditions, any new substrate should proof its durability under conditions of a daily farming routine. This would include mechanical thinning, harvesting processes, and tests on the reusability of the material.

The technical realization and the implications of aquaculture technical requirements on design and construction of the grounding construction of offshore



**Aquaculture and Renewable Energy Systems, Integration of. Figure 7**

Modeling of potential attachment points for the combination of longline connections to a tripod foundation. (a) Displays alternative connection points, (b) shows the generation of representative loads on the wind-energy installation, including vibrations, (c) shows the respective tripod foundation for offshore use in depths of about 20–50 m, and (d) shows the development of a static model (3–5 MW class) [47]

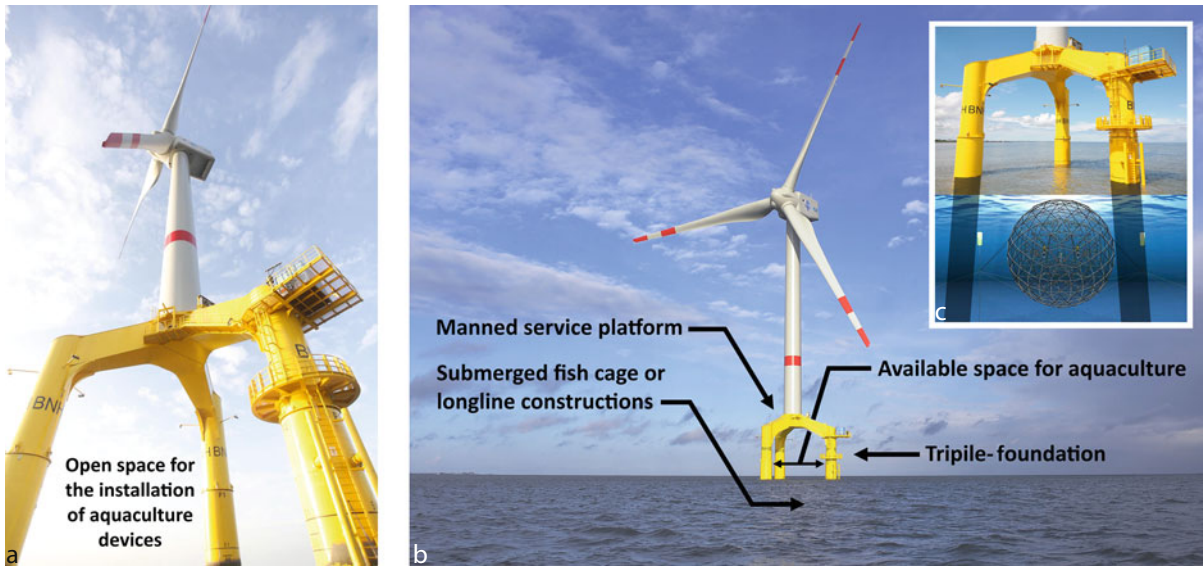
wind turbines were considered. So far, modeling and experimental validation of a submerged 50 m longline aquaculture construction mounted between two steel piles, 17 nautical miles off the coast, show significant forces of up to 90 kN (equivalent to 9 t) induced by waves of up to 1.8 m significant wave height and tidal currents of up to 1.0 m/s [81]. Given the high-energy environment in the North Sea and the non-linear relationship between water movement and its resulting forces, even higher mechanical loads are to be expected within the life cycle of such an arrangement. These must be taken into account when developing techniques for larger-scale offshore cultivation within wind farms.

Finally, a new cage design project has been initiated, where it will be investigated whether aquaculture of fish in between a tripile construction below a windmill has the potential to enlarge the diversity of candidates to be grown offshore (next to bivalve and seaweed) as well as widening the potential of offshore farming within wind farms. First insights are shown in Fig. 8 [82, 83].

### Management and Institutional Considerations

From a spatial planning perspective, the ocean space in the Exclusive Economic Zones (EEZ) cannot be considered any more as “commons” in the sense of Ostrom





**Aquaculture and Renewable Energy Systems, Integration of. Figure 8**

Tripile construction for the secondary use for fish cages. (a) Shows the open space within a tripile foundation to be used for aquaculture purposes, (b) displays a lateral view of the Bard Windmill and the access to the fish cage, and (c) is a photo animation and gives an idea how a fish farm, such as an aquapod, could be moored below [82, 83]

et al. [84] wherein individuals or groups have the right to freely consume and return any kind of resources. As a matter of fact, the “tragedy of the commons” situation Hardin described in 1968 [85], has already been reached for most of the oceans today. Offshore waters are in a process of transition, revealing diverse and heterogenic interests in marine resources. For instance, the development of offshore renewable-energy systems is an international priority driven by the need to reduce the dependence on fossil fuels and decrease human impacts on the global climate regime. Simultaneously, the demand for high-quality seafood is accelerating globally. This leads to an increased complexity and thus to limitations in developing and managing the different and often spatially overlapping maritime activities independently of one another. The upcoming new utilization patterns of the German North Sea, such as wind farms, but also Marine Protected Areas (MPAs) reported to the EU commission in Brussels for the “European Natura 2000 network,” reveal a trend toward the development of permanent constructs. Both are examples for new forms of use with a high spatial demand [86]. Not all uses are compatible with each other and user conflicts with

existing activities, such as fisheries, maritime traffic, or military missions are preordained. The planned large-scale offshore wind farms as well as designated MPAs are prime examples for the development of lasting marine structures that take up a surface area of several square kilometers each [55].

At the same time, the increasing demand for high-quality foods worldwide accelerates the development of marine aquaculture. This potential newcomer can be expected to become an additional competitor in offshore waters [87], contributing to the increase in spatial competition and complexity in the ocean [20]. Conflicts among the respective user groups are inevitable. The growing competition for space represents a major challenge for further developing or even maintaining all forms of marine aquaculture, as well as freshwater fish farming. However, area choice is crucial and spatial planning has a key role to play in providing guidance and reliable data for the location of an economic activity, giving certainty to investors, avoiding conflicts, and finding synergies between activities and environments with the ultimate aim of sustainable development [88]. The inclusion of all



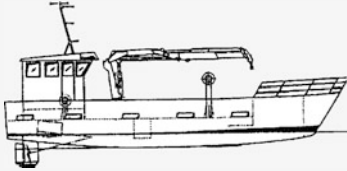
stakeholders in this process to find synergies in the open ocean is crucial.

Ongoing multidisciplinary social-science research in Europe shows that it is feasible to establish spatially efficient and effective wind farm–mariculture co-management regimes. A window-of-opportunity has opened as both groups have realized that they may benefit through the integration of operation and maintenance (O&M) activities vis á vis gaining support in collaborative action by the current impetus of the new EU Maritime Policy. The operation and maintenance of any offshore installation is a major challenge due to restricted logistics and accessibility, forming a large part of the overall costs. A five-to-ten time more expensive scale of operation and more difficult logistics for

maintenance and/or harvesting compared to nearshore or onshore sites have to be taken into account [89–91]. Experiences with existing wind farms and mariculture sites off the coast show that work at the sea is not only significantly more cost-intensive, but also more time consuming than on land [92].

There are certain rights and duties involved if prospective spatial and organizational interaction of O&M activities of offshore wind turbines and mariculture installations are to be combined [20]. Different values, perspectives, and demands of the stakeholder groups need to be harmonized [93]. So far, disagreements on the distribution of entitlements to benefits and profits between the different stakeholder groups can be observed (Table 1). The two potential adopters of

**Aquaculture and Renewable Energy Systems, Integration of. Table 1** Offers, needs, and constraints characteristics of mariculture operators and offshore wind farmers concerning O&M activities. Interrelated aspects between the two actor groups are indicated in bold (modified after [86])

| Characteristics | Actor groups                                                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                 | Wind farmers                                                                                                                                                                                                                                                                                                                                                                                                | Mariculture operators                                                                                                                                                                                                                                                                                                                                                                          |
|                 |                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                            |
| Offers          | <ul style="list-style-type: none"> <li>• Fixed offshore infrastructure</li> <li>• Logistic platform</li> <li>• Financial support (EEG amendment)</li> </ul>                                                                                                                                                                                                                                                 | <ul style="list-style-type: none"> <li>• Upgradeable sea-going vessels</li> <li>• Offshore mentality</li> <li>• Offshore skills and experience</li> </ul>                                                                                                                                                                                                                                      |
| Needs           | <ul style="list-style-type: none"> <li>• Specialization of equipment (construction vs hire; “marinization” of onshore equipment)</li> <li>• Specialization of personnel</li> <li>• Sea-going vessels</li> <li>• Service demands (man-hours)</li> <li>• Suitable O&amp;M pattern (corrective vs preventive maintenance)</li> <li>• Suitable O&amp;M pattern (opportunity vs periodic maintenance)</li> </ul> | <ul style="list-style-type: none"> <li>• Specialization of equipment (construction vs alteration of existing oil industry/fishery vessels)</li> <li>• Specialization of personnel</li> <li>• Fixed offshore infrastructure</li> <li>• Technical and logistic support</li> <li>• Service demands (man-hours)</li> <li>• Offshore skills and experience</li> <li>• Offshore mentality</li> </ul> |
| Constraints     | <ul style="list-style-type: none"> <li>• Operation costs</li> <li>• Technical challenges</li> <li>• Distance to farm site</li> <li>• Available working days (estimated 100/year)</li> <li>• Difficult logistics for O&amp;M</li> <li>• Reliability of offshore wind turbines</li> </ul>                                                                                                                     | <ul style="list-style-type: none"> <li>• Access to farm site (uncertain regulatory and permit requirements)</li> <li>• Distance to farm site</li> <li>• Available working days (estimated 30–100/year)</li> <li>• Difficult logistics for maintenance and harvesting</li> <li>• Reliability of culturing devices</li> </ul>                                                                    |

such a multi-use scheme illuminate different sets of skills and capacities in terms of offers, needs, and constraints characteristics. These are vital resources, which provide the basis for forming any sustainable offshore co-management arrangements [85]. Hereby, a fair negotiation and bargaining process is the most essential component to effectively orchestrate co-management of offshore wind farmers and future mariculture operators, such as mussel harvesters. The latter already dispose vital skills and experiences for working in the open sea. Still, working methods have to be adjusted to the offshore culture production mode.

If such an offshore co-management is considered as a network activity between private actors, such as wind farmers or mariculture operators/fishermen and public authorities, one of its basic characteristics is the fact that a third party can coordinate the activities of formally separated parties [94]. Ways and means have to be developed that balance the respective interests of dominant and politically supported wind farming participants with small-scale entrepreneurial mariculturists. The key question is how institutional arrangements could act as “boundary organizations” [95] in an offshore co-management process. Such a process is more likely to develop and succeed if an interface management that acts as moderator, disclosing the interests of the actor groups and offering possibilities for concerted action, guides it. With respect to the decision-making arrangements at the three levels (operational level, organizational level, and legislative level), the interface management would thus help to determine the rules for interaction among the actor groups and state authorities at the organizational level. Besides, it would facilitate organizing and decision making of the day-to-day activities at the operational level. However, to authorize and legitimize new co-management arrangements for interacting offshore O&M activities, new policies must be developed or existing laws amended. Following a dynamic process of forming new institutional structures, the establishment of a communication arena may (a) support a common understanding of the entire co-management process, (b) provide the overall framing for an improved communication among the participating actor groups, (c) increase the level of trust among the actor groups, and (d) promote sustainability and efficiency in times of scarcity of spatial resources [85]. However, top-down induced management schemes

by, e.g., the national government, hold a high potential for failure. Involving the relevant actors improves the social acceptability of innovative concepts and their applicability [96]. Consequently, it appears that for developing and implementing a wind farm-mariculture multiple-use concept, co-management, such as that described by Carlsson and Berkes [94], should ideally be carried out with the participation of different actors that typically try to find ways to learn from their actions and adapt the behavior to the consequences of their own and other’s actions. This must be supported by the relevant authorities at all levels and must find its way into the legislative framework at the EU and national level.

On EU level, the issue of access to space for maritime activities, including aquaculture, has been recognized in several communications over the past years, e.g., in 2007 pertaining to the Integrated Maritime [97] Policy or in 2009 concerning a new impetus for sustainable aquaculture in Europe. In the latter, all Member States are asked to develop marine spatial planning systems, in which they fully recognize the strategic importance of aquaculture. This Strategy also aims at providing EU leadership and guidance to both stakeholders and administrations to ensure consistency and clarity in designing the necessary policies for the future sustainable development of European aquaculture. In this context, a partnership between public authorities and interested parties at EU, national, and local level play a crucial role.

Hence, European aquaculture should benefit from an improved framework for governance; however, it is stressed that the national authorities have a primary role in shaping aquaculture development in their territory. While in some countries aquaculture is defined and regulated under the agricultural laws, in other countries regulations are dispersed, and consequently the responsibilities are in the hand of several agencies with no clearly defined lead agency. So far, a number of important challenges that limit the development of European aquaculture directly depend on policies and actions taken at national or regional level. A bottom-up approach is therefore needed so that the public authorities can establish an appropriate framework for the vision of multiple-use of offshore areas to become operational. A participatory approach contributes to lifting bottlenecks in national legislation.

This framework needs to be transparent, consistent and cost-effective in order to allow the industry to realize its potential. Unless these and other regulative issues pertaining to multi-use offshore conditions remain unresolved, an establishment of offshore co-management arrangements may be very difficult. Still, the current lack of legislation in the EEZ holds the potential to implement concerted innovative concepts of offshore constructions and thereon-interacting activities.

However, several restrictions are still needed to be resolved. Questions pertaining to access rights within the wind farm area have to be deciphered. So far, approved offshore wind farm territories in, e.g., the German EEZ, are designated as restricted area, prohibiting any kind of public access [98, 99]. However, conveying access rights to a second party is inevitable if wind farm O&M is performed by a commissioned subcontractor and not by the licensee itself. In a wind farm–mariculture multiple-use arena, the same access and user rights have to be guaranteed to a mariculture operator who enters the territory for purposes not related to wind farming but for maintaining culturing devices and harvesting procedures. In this case, precise positioning of aquaculture installations within the wind farm territory as well as access lanes for both parties have to be specified.

In addition, the question of harmonizing the tenure or duration of a lease for offshore resources has to be tackled. If there is, i.e., significant discrepancy in the length of lease tenures between the two uses to be combined, the resource users may not be inclined to create long-term co-management arrangements. Furthermore, cooperative management structures also benefit if the leasing process was combined and/or effectively coordinated, since it facilitates, i.e., integrating O&M within a co-management scheme once the projects are operational.

Yet, in order to define the functional structure of such a co-management regime in detail, reliable outcomes on economic and technical integration prospects of a joint wind farm–mariculture venture have to be produced. The latter is a major research demand, which was voiced by most of the interview partners along the North Sea coast so far [93]. Cumulative impacts of different economic sectors, such as offshore wind farms and mariculture need to be addressed, which provides an opportunity

to create synergies between different industrial sectors prior to their installation.

### Outcome of Economic Studies

The *Feasibility Study* [13, 24] provided a general overview on market prices, market demands, classification of candidate species as high-value products, and the cost of some infrastructure. The study showed possible market value of offshore aquaculture products in comparison to the performance of existing conventionally operated farms in coastal waters.

Basic data for offshore mussel cultivation in close vicinity to a designated offshore wind farm in the open sea of the German Bight were compiled. It contained different case-scenario calculations to illustrate the impact of changing parameter values on overall profitability or non-profitability of this activity. Primary focus was placed on the production of consumer mussels, but seed mussel cultivation is also taken into consideration. This study concluded with providing some recommendations on how favorable terms or actions could further improve profitability of offshore mussel cultivation. Results intended to shed some light on business management topics that future offshore mariculture operators should follow in order to be efficient [100].

Nonetheless, the economics of a joint offshore wind farm–mariculture utilization scheme still remain to be evaluated in more detail.

### Future Directions

By setting higher value on an inclusion of stakeholder knowledge and opinions, the initiation of the *Coastal Futures Project* resulted in a stronger focus on the practicability of multifunctional use of offshore areas. It can be shown that such innovative new concepts are highly complex and interdependent. First, results indicate that secure technical and economic feasibility appears to be a basic prerequisite to assure that both offshore wind farm operators and aquaculturists will support the multi-use concept, especially as far as the management of joint activities is concerned.

This suggests that as soon as technical and economic aspects are evaluated in more detail, it is important to initialize a comprehensive communication program to provide information to the key public and private actor groups (stakeholders). Furthermore, effective and continuous participation of all

stakeholders on all levels from the very beginning of the multi-use approach must be ensured. This supports the orchestration of scientific and local user knowledge in an overall approach to combine different offshore uses. In addition, it contributes to adding a joint wind farm–mariculture venture to their future portfolio.

More detailed data are needed to calculate the economic potentials and risks of a co-used wind farm area for the production of seafood. Apart from the principal feasibility of an area as an aquaculture site, growth rates and product quality must be predictable. First, results on blue mussels from test areas show that highest product qualities can be expected from testing areas offshore. A proven product quality ensures higher market prizes, should compensate for higher investment costs for the culture systems, and help to install a functioning offshore aquaculture system in the German Bight.

Generally, science for open ocean aquaculture needs a transformative moment. It seems necessary to learn the skills to interact constructively with different scientific disciplines and different stakeholders. This will require a new science for managed marine seascapes [101]. Creating a system biology paradigm in ecosystem science and aquaculture will require a multidisciplinary input, with scientific interactions not just at the margins of each discipline, but focused collaboratively on the realization of a vision of multifunctional, spatially effective, and sustainable use of ocean space. This will require new kinds of scientists (with new kinds of career structures) who are trained to work in multidisciplinary teams. The need for such training is now widely recognized and is reflected in the emerging curricula's of many new MSc courses.

It is mandatory to discover what to do, at what scale, in what modality – engineering, farming, legislation, social organization, economic initiatives, etc. – and how to do it. Since the activities in the ocean realm are concerted in integration, future activities must also be integrated over all these modalities.

## Bibliography

1. FAO/FIDI (1989) Aquaculture production (1984–1986) Food and Agriculture Organisation of the United Nations. Rome Fisheries Circular 815, 106
2. FAO (1997) Aquaculture development. Technical guidelines for responsible fisheries. Food and Agriculture Organisation of the United Nations, Rome, pp 5–40
3. EU Commission (2000) On integrated coastal zone management: a strategy for Europe. Communication from the Commission to the Council and the European Parliament, COM (2000) 547 final
4. Garvine R, Kempton W (2008) Assessing the wind field over the continental shelf as a resource for electric power. *J Mar Res* 66(6):751–773
5. Pauly D, Christensen V, Guénette S, Pitcher TJ, Sumaila RU, Walters CJ, Watson R, Zeller D (2002) Towards sustainability in world fisheries. *Nature* 418:689–695
6. FAO (2004) The state of world fisheries and aquaculture-2004 (SOFIA). Fisheries Department. Food and Agriculture Organisation of the United Nations, Rome, Italy
7. FAO (2009) The State of World Fisheries and Aquaculture (SOFIA) 2008, FAO Fisheries and Aquaculture Department. Food and Agriculture Organisation of the United Nations
8. Buck BH (2007a) Experimental trials on the feasibility of offshore seed production of the mussels *Mytilus edulis* in the German Bight: installation, technical requirements and environmental conditions. *Helgol Mar Res* 61:87–101
9. Buck BH (2007b) Marikultur als Co-Nutzung in Offshore-Windparks: Status Quo, Probleme und Perspektiven, Meeresumwelt-Symposium 2006: 16. Symposium, 13. bis 14. Juni 2006, CCH – Congress Center Hamburg/Bundesamt für Seeschifffahrt und Hydrographie in Zusammenarbeit mit dem Bundesumweltamt im Auftrag des Bundesministeriums für Umwelt, Naturschutz und Reaktorsicherheit, pp 167–179
10. FAO (2006) Fishery information, data and statistics unit. Aquaculture production: values 1984–2004. FISHSTAT Plus – universal software for fishery statistical time series. Food and Agriculture Organization of the United Nations, Rome, Italy
11. Burbridge P, Hendrick V, Roth E, Rosenthal H (2001) Social and economic policy issues relevant to marine aquaculture. *J Appl Ichthyol* 17:194–206
12. Rawski TG, Xiao W (2001) Roundtable on Chinese economic statistics: introduction. *Chin Econ Rev* 12:298–302
13. Naylor RL, Goldburg RJ, Primavera JH, Kautsky N, Beveridge MCM, Clay J, Folke C, Lubchenco J, Mooney H, Troell M (2000) Effect of aquaculture on world fish supplies. *Nature* 405:1017–1024
14. Roth E, Ackefors H, Asche F, Balnath C, Black E, Black K, Bogen A, Browdy C, Burbridge P, Castell JD, Chamberlain G, Dabrowski K, Davies I, Dosdat A, Eleftheriou A, Ervik A, Gordin H, Heinig CS, Hilge V, Karakassis I, Kuhlmann H, Landry T, von Lukowicz M, McGlade J, Price A, Rheault RB, Rosenthal H, Saint-Paul U, Sandifer PA, Saroglia M, Silvert W, SteVens W, Soto D, Varadi L, Verreth J, Verdegem M, Waller U (2002) An intellectual injustice to aquaculture development: a response to the review article on “Effect of aquaculture on world fish supplies”, Report of the ICES working group on environmental interactions of mariculture, F:04 REF ACME, Annex 4, pp 83–89
15. Buck BH (2004) Farming in a high energy environment: potentials and constraints of sustainable offshore aquaculture in the

- German Bight (North Sea). Dissertation. University of Bremen, Germany
16. Davis A, Bailey C (1996) Common in custom, uncommon in advantage: common property, local elites, and alternative approaches to fisheries management. *Soc Nat Resources* 9:251–265
  17. Adger WN, Luttrell C (2000) Property rights and the utilisation of wetlands. *Ecol Econ* 35:75–89
  18. Rana KJ (1997) Aquatic environments and use of species sub-groups. In: Review of the state of world aquaculture. FAO Fisheries Circular FIRI/C886 (Rev. 1). Food and Agriculture Organisation (FAO), Rome, Italy, pp 7–16
  19. Lee PG, Turk PE (1998) Overview of a modern, shore-based hatchery for offshore mariculture support. In: Stickney RR (ed) *Joining forces with industry – open ocean aquaculture. Proceedings of the third annual international conference*, May 10–15, Texas Sea Grant College Program, Corpus Christi, pp 87–102
  20. Krause G, Buck BH, Rosenthal H (2003) Multifunctional use and environmental regulations: potentials in the offshore aquaculture development in Germany. *Proceedings of the multidisciplinary scientific conference on sustainable coastal zone management "Rights and duties in the coastal zone"*, 12–14 June 2003. Stockholm, Sweden
  21. Buck BH, Krause G, Rosenthal H, Smetacek V (2003) Aquaculture and environmental regulations: the German situation within the North Sea. In: Kirchner A (ed) *International marine environmental law: institutions, implementation and innovation*, vol. 64. International Environmental. Law and Policies Series of Kluwer Law International, The Hague, pp 211–229
  22. Firestone J, Kempton W, Krueger A, Loper CE (2004) Regulating offshore wind power and aquaculture: messages from Land and sea. *Cornell J Law Public Policy* 14(1):71–111
  23. Langan R (2009) Co-location of offshore energy and seafood production: potential synergies, compatibilities and conflicts. Conference paper, the ecology of marine wind farms: perspectives on impact mitigation, siting, and future uses, 8th annual Ronald C. Baird Sea Grant Science Symposium, November 2–4 2009, Newport, Rhode Island
  24. BMU (2006) Integriertes Küstenzonenmanagement in Deutschland (IKZM). Nationale Strategie mit Bestandsaufnahme nach der EU-Empfehlung 2002/413/EG. Bundesregierung für Umwelt, Naturschutz und Reaktorsicherheit (Federal Ministry for the Environment, Nature Conservation and Nuclear Safety)
  25. Kannen A (2004) Holistic systems analysis for ICZM: the Coastal Futures approach. In: Schernewski G, Dolch T (eds) *Geographie der Meere und Küsten. Coastline reports*, vol 1, pp 171–181
  26. Buck BH (2002) Open Ocean Aquaculture und Offshore-Windparks: Eine Machbarkeitsstudie über die multifunktionale Nutzung von Offshore-Windparks und Offshore-Marikultur im Raum Nordsee. Reports on Polar and marine research. Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, pp 412–252
  27. Ryan J (2005) Offshore aquaculture – do we need it, and why is it taking so long? International Salmon Farmers Association (Ireland). Expert workshop on sustainable aquaculture, DG JRC European Commission, Institute for Prospective Technological Studies, 17–18 January 2005. Seville, Spain
  28. Dahle LA, DePauw N, Joyce J (1991) Offshore aquaculture technology – possibilities and limitations. *Aquacult Environ* 14:83–84
  29. Eleftheriou M (1997) *Aqualex: a glossary of aquaculture terms*. Wiley-Praxis series in aquaculture and fisheries, Aqualex Multimedia Consortium Ltd
  30. Chopin T (2009) Let's not move just the fish to the open ocean... an integrated multi-trophic aquaculture approach should not be an afterthought for 2050. Conference paper, the ecology of marine wind farms: perspectives on impact mitigation, siting, and future uses, 8th annual Ronald C. Baird Sea Grant Science Symposium, November 2–4, 2009, Newport, Rhode Island
  31. Dena (2007) Renewable energies: Germany's first test field for offshore wind energy. Dena fact sheet. German Energy Agency, Berlin
  32. Tiedemann A (2003) Windenergieparke im Meer – Perspektiven für den umweltverträglichen Einstieg in eine neue Großtechnologie. In: Lozán J, Rachor E, Reise K, Sündermann J, Westernhagen HV (eds) *Warnsignale aus Nordsee & Wattenmeer: Eine aktuelle Umweltbilanz. Wissenschaftliche Auswertungen*, Hamburg, pp 142–148
  33. BMU/SOW (2007) Offshore wind power deployment in Germany. In: Rehfeldt K, Paschedag U, Bömer J (eds) *Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU) and Stiftung der Deutschen Wirtschaft zur Nutzung und Erforschung der Windenergie auf See/Offshore Wind Energy Foundation (SOW)*. Scholz-Druck GmbH, Dortmund, pp 31
  34. Gierloff-Emden HGR (2002) Wandel der Umwelt der See- und Küstenlandschaft der Nordsee durch Nutzung von Windenergie. *Mitteilungen der Österreichischen Geographischen Gesellschaft* 144:219–226
  35. Dahlke V (2002) Genehmigungsverfahren von Offshore-Windenergieanlagen nach der Seeanlagenverordnung. *Natur und Recht* 24:472–479
  36. BWE (2008) Datenblatt Windenergie 2007. Bundesverband WindEnergie
  37. GWEC (2007) Continuing boom in wind energy – 20 GW of new capacity in 2007. Press release 15/02 200. Global Wind Energy Council
  38. BSH (2010) Wind farms. Bundesamt für Seeschifffahrt und Hydrographie. Federal Maritime and Hydrographic Agency, Hamburg/Rostock
  39. Ventus A (2010) Der erste deutsche Offshore-Windpark (first german offshore wind farm). Deutsche Offshore-Testfeld und Infrastruktur. GmbH & Co. KG, Oldenburg
  40. BWE (2007) Offshore Windenergie. BWE Hintergrundpapiere. Bundesverband WindEnergie



41. Wirtz KW, Tol RSJ, Hooss KG (2003) Mythos "Offene See": Nutzungskonflikte im Meeresraum. In: Lozan L et al (eds) Warnsignale aus Nordsee & Wattenmeer. Eine aktuelle Umweltbilanz. Wissenschaftliche Auswertungen, Hamburg, pp 157–160
42. Buck BH, Krause G, Rosenthal H (2004) Multifunctional use, environmental regulations and the prospect of offshore co-management: potential for and constraints to extensive open ocean aquaculture development within wind farms in Germany. *Ocean Coast Manage* 47:95–122
43. Jentoft S (2000) Co-managing the coastal zone: is the task too complex? *Ocean Coast Manage* 43:527–535
44. Michler-Cieluch T (2009) Co-Management processes in integrated coastal management: the case of integrating marine aquaculture in offshore wind farms. PhD thesis, University of Hamburg, Germany
45. Führböter A, Dette HH (1983) Wasserstände, Wind und Seegang im Seegebiet um Helgoland in den Jahren von 1969 bis 1983. Leichtweiß-Institut für Wasserbau der TU Braunschweig, Bericht Nr. 577
46. Becker G, Dick S, Dippner J (1992) Hydrography of the German Bight. *Mar Ecol Prog Ser* 91:9–18
47. Buck BH, Berg-Pollack A, Assheuer J, Zielinski O, Kassen D (2006) Technical realization of extensive aquaculture constructions in offshore wind farms: consideration of the mechanical loads. In: Proceedings of the 25th international conference on offshore mechanics and Arctic engineering, OMAE 2006: presented at the 25th International conference on offshore mechanics and Arctic engineering, 4–9 June 2006, Hamburg, Germany/American Society of Mechanical Engineers, pp 1–7
48. Takayanagi K (1998) Water quality guidelines for aquaculture: an example in Japan. In: Howell WH, Keller BJ, Park PK, McVey JP, Takayanagi K, Uekita Y (eds) Nutrition and technical development of aquaculture. Proceedings of the 26th US–Japan aquaculture symposium, Durham/New Hampshire/USA September 16–18, 1997. UJNR technical report No. 26, Durham, University of New Hampshire Sea Grant Program, pp 247–254
49. BSH (2006) MURSYS – marine environment reporting system information from the North Sea and Baltic Sea. Bundesamt für Seeschifffahrt und Hydrographie (BSH), Hamburg
50. Stickney RR (1998) Joining forces with industry – open ocean aquaculture. Proceedings of the 3rd annual international conference, May 10–15, Corpus Christi. TAMU-SG-99–103, Corpus Christi, Texas Sea Grant College Program
51. Bridger CJ, Costa-Pierce BA (2003) Open ocean aquaculture: from research to commercial reality. The World Aquaculture Society, Baton Rouge
52. Michler-Cieluch T, Kodeih S (2008) Mussel and seaweed cultivation in offshore wind farms: an opinion survey. *Coast Manage* 36(4):392–411
53. Buck BH, Buchholz CM (2004) The offshore-ring: a new system design for the open ocean aquaculture of macroalgae. *J Appl Phycol* 16:355–368
54. Buck BH, Walter U, Liebezeit G (2011) Larval occurrence and settlement in the German Bight: A trial to estimate potentials for *Mytilus edulis* culture in offshore areas, Helgoland Marine Research (in press)
55. BSH (2009) Bundesamt für Seeschifffahrt und Hydrographie – BSH (Federal Maritime and Hydrographic Agency), CONTIS-Geodata. Hamburg/Rostock
56. Hickman RW (1992) Mussel cultivation. In: Gosling E (ed) The mussel *Mytilus*: ecology, physiology, genetics and culture, vol 25, Development in aquaculture and fisheries science. Elsevier, Amsterdam, pp 465–510
57. Critchley AT, Ohno M, Largo DB (2006) World seaweed resources. An authoritative reference system. ETI Bioinformatics. DVDROM
58. Tseng CK (1989) *Laminaria* mariculture in China. In: Doty MS, Caddy JF, Santelices B (eds) Case studies of seven commercial seaweed resources. FAO Fisheries Technical Paper – 281. Food and Agriculture Organization of the United Nations (FAO), Rome
59. Strathmann MF (1987) Reproduction and development of marine invertebrates of the northern Pacific Coast. University of Washington Press, Seattle, pp 670
60. Pulfrich A (1997) Seasonal variation in the occurrence of planktic bivalve larvae in the Schleswig-Holstein Wadden Sea. *Helgol Wiss Meer* 51:23–39
61. Walter U, Liebezeit G (2001) Abschlußbericht des Projektes: Nachhaltige Miesmuschel-Anzucht im niedersächsischen Wattenmeer durch die Besiedlung natürlicher und künstlicher Substrate. Forschungszentrum Terramare, pp 1–97
62. Walter U, Liebezeit G (2003) Efficiency of the blue mussel (*Mytilus edulis*) spat collectors in highly dynamic tidal environments of the Lower Saxonian coast (Southern North Sea). *Biomol Eng* 20:407–411
63. Scarratt D (1993) A handbook of northern mussel culture. Island Press, Montague, pp 167
64. Gosling E (2003) Bivalve molluscs: biology, ecology and culture. Blackwell Publishing/MPG Books, Bodmin/Cornwall, pp 443
65. Brenner M (2009) Site selection criteria and technical requirements for the offshore cultivation of blue mussels (*Mytilus edulis* L.). PhD thesis, School of Engineering and Science, Jacobs University, Bremen, pp 151
66. Langan R, Horton F (2003) Design, operation and economics of submerged longline mussel culture in the open ocean. *Bull Aquacult Assoc Can* 103:11–20
67. Neushul M, Harger BWW (1985) Studies of biomass yield from a near-shore macroalgal test farm. *J Solar Energy Eng* 107:93–96
68. Neushul M, Benson J, Harger BWW, Charters AC (1992) Macroalgal farming in the sea: water motion and nitrate uptake. *J Appl Phycol* 4:255–265
69. Turner R (2001) Offshore mariculture: site evaluation. In: Muir J, Basurco B (eds) Options méditerranéennes – Mediterranean offshore mariculture. Etudes et recherches, Serie B, Numéro 30, Zaragoza, CIHEAM, INO Reproducciones, pp 141–157

70. Pérez OM, Telfer TC, Ross LG (2003) On the calculation of wave climate for offshore cage culture site selection: a case study in Tenerife (Canary Islands). *Aquacult Eng* 29:1–21
71. Dalton R (2004) Fishing for trouble. *Nature* 431:502–504
72. Naylor R, Burke M (2005) Aquaculture and ocean resources: raising tigers of the sea. *Annu Rev Environ Res* 30:185–218
73. Allen JI, Moore MN (2004) Environmental prognostics: is the current use of biomarkers appropriate for environmental risk evaluation. *Mar Environ Res* 58:227–232
74. Buck BH, Walter U, Rosenthal H, Neudecker T (2006) The development of mollusc farming in Germany: past, present and future. *World Aquaculture* 37(2):6–11 & 66–69
75. Buck BH, Thieltges D, Walter U, Nehls G, Rosenthal H (2005) Inshore–offshore comparison of parasite infestation in *Mytilus edulis*: implications for open ocean aquaculture. *J Appl Ichthyol* 21:107–113
76. Brenner M, Buck BH, Koehler A (2007) New concept combines offshore wind farms, mussel cultivation. *Global Aquacult Advocate* 10:79–81
77. Buck BH, Buchholz CM (2005) Response of offshore cultivated *Laminaria saccharina* to hydrodynamic forcing in the North Sea. *Aquaculture* 250:674–691
78. Manefeld T (2006) Ansiedlungspotential von Miesmuschellarven (*Mytilus edulis*) an verschiedenen Substraten freihängender Brutkollektoren. Thesis, University of Applied Sciences Bremerhaven, Germany
79. Voss D (2006) Parasitierungsgrad der Miesmuschel (*Mytilus edulis*) in Abhängigkeit zur Tiefenzonierung und Entfernung zur Küste. Thesis, University of Applied Sciences Bremerhaven, Germany
80. Brenner M, Buck BH (2010) Attachment properties of blue mussel (*Mytilus edulis* L.) byssus threads on culture-based artificial collector substrates. *Aquacult Eng* 42:128–139
81. Zielinski O, Assheuer J, Berg-Pollack A, Buck BH, Geisen M, Henkel R, Kassen D (2006) Assessment of mechanical loads and environmental conditions for extensive aquaculture constructions within offshore wind farms: First results from the AquaLast study site. Proceedings of DEWEK 2006: presented at the DEWEK 2006, 22–23 November 2006, Bremen, Germany, pp 1–4
82. Bard (2010) Offshore wind farms. Bard Engineering GmbH. Emden, Germany
83. OFT (2010) *Aquapod* – a submersible fish cage. Ocean Farm Technologies, Seasmont
84. Ostrom E, Burger J, Field CB, Norgaard RB, Policansky D (1999) Revisiting the commons: local lessons, global challenges. *Science* 284(4512):278–282
85. Hardin G (1968) The tragedy of the commons. *Science*, 162. Reprinted in Dryzek JS, Schlosberg D (1999) *Debating the Earth: the environmental politics reader*. Oxford University Press, New York
86. Michler-Cieluch T, Krause G, Buck BH (2009) Reflections on integrating operation and maintenance activities of offshore wind farms and mariculture. *Ocean Coast Manage* 52(1):57–68
87. Buck BH, Krause G, Michler-Cieluch T, Brenner M, Buchholz CM, Busch JA, Fisch R, Geisen M, Zielinski O (2008) Meeting the quest for spatial efficiency: progress and prospects of extensive aquaculture within offshore wind farms. *Helgol Mar Res* 62:269–281
88. EU Commission (2009) Building a sustainable future for aquaculture. A new impetus for the Strategy for the Sustainable Development of European Aquaculture. The European Parliament and the Council, COM(2009) 162 final
89. McVey JP (1996) Overview of offshore aquaculture. In: Polk M (ed) *Open ocean aquaculture. Proceedings of an international conference, May 8–10, Portland, Maine*. New Hampshire/Maine Sea Grant College Program, pp 13–18
90. Braginton-Smith B, Messier RH (1997) Design concepts for integration of open ocean aquaculture and Osprey TM Technology. In: Howell WH, Keller BJ, Park PK, McVey JP, Takayanagi K, Uekita Y (eds) *Proceedings of the twenty-sixth US–Japan aquaculture symposium, Durham, New Hampshire*. UJNR Technical Report No. 26. University of New Hampshire Sea Grant Program, Durham, pp 239–245
91. Bussel van GJW, Zaaijer MB (2007) Reliability, availability and maintenance aspects of large-scale offshore wind farms, a concepts study. In: *Proceedings of MAREC 2001*. Delft University of Technology, The Netherlands, pp 119–126
92. Musial W, Butterfield S, Ram B (2006) Energy from offshore wind. In: *Proceedings of the offshore technology conference (OTC), 1–4 May, Houston, 11 pp*
93. Michler-Cieluch T, Krause G (2008) Perceived concerns and possible management strategies for governing ‘wind farm–mariculture integration’. *Marine Policy* 32(6): 1013–1022
94. Carlsson L, Berkes F (2005) Co-management: concepts and methodological implications. *J Environ Manage* 75: 65–76
95. Cash DW, Moser SC (2000) Linking global and local scales: designing dynamic assessment and management processes. *Global Environ Change* 10:109–120
96. Heinelt H (2002) Achieving sustainable and innovative policies through participatory governance in a multi-level context. In: Heinelt H, Getimis P, Kafkalas G, Smith R, Swyngedouw E (eds) *Participatory governance in multi-level context*. Leske + Budrich, Opladen, pp 17–32
97. EU Commission (2007) An Integrated Maritime Policy for the European Union. Commission Staff Working Document, Accompanying document to the Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions, SEC (2007) 1278
98. Dietz T, Ostrom E, Stern PC (2003) The struggle to govern the commons. *Science* 302(5652):1907–1912
99. Steins NA, Edwards VM (1999) Synthesis: platforms for collective action in multiple-use common-pool resources. *Agric Hum Values* 16(3):309–315

100. Buck BH, Ebeling M, Michler-Cieluch T (2010) Mussel cultivation as a co-use in offshore wind farms: potential and economic feasibility. *Aquacult Econ Manage* 14(4): 1365–7305
101. Bradbury RH, Seymour RM (2009) Coral reef science and the new commons. *Coral Reefs* 28:831–837

## Aquaculture, Ecological

BARRY A. COSTA-PIERCE

Graduate School of Oceanography, Rhode Island Sea, Grant College Program, University of Rhode Island, URI Bay Campus, Narragansett, RI, USA

### Article Outline

Glossary  
Definition of Ecological Aquaculture  
Introduction  
Key Principles  
The FAO Ecological Approach to Aquaculture (EAA)  
Applying an Ecological Aquaculture Approach at Different Scales of Society  
Social Ecology of Aquaculture  
An Ecological Aquaculture Strategy for the “Triple Bottom Line”  
Future Directions  
Bibliography

### Glossary

- Biofloc** A mixture of detritus, bacteria, and other microscopic organisms that aggregates in flocs, which are used for controlling water quality and enhancing the delivery of natural foods to omnivorous species in aquaculture.
- Ecosystem** An area of the natural environment in which the structure and functions of the physical (rocks, soil, etc.) and natural (all living organisms) environments are considered together in interacting food webs.
- Escapees** The unintended releases of cultured organisms from captivity into the wild.
- Polyculture** The practice of making compatible the culture of multiple species in the same physical space by stocking or planting organisms having different food, spatial, or temporal niches.

**Resilience** The ability of a natural or aquaculture system to absorb abrupt changes or disturbances without collapsing. A resilient aquaculture ecosystem can withstand physical and economic shocks and rebuild itself.

**Stewardship** An ethic that engages all affected stakeholders in the cooperative planning and management of the environmental quality to prevent degradation and facilitate recovery in the interest of long-term sustainability.

**Watershed** An area of land where all of the water that is under it or drains off of it goes into the same place.

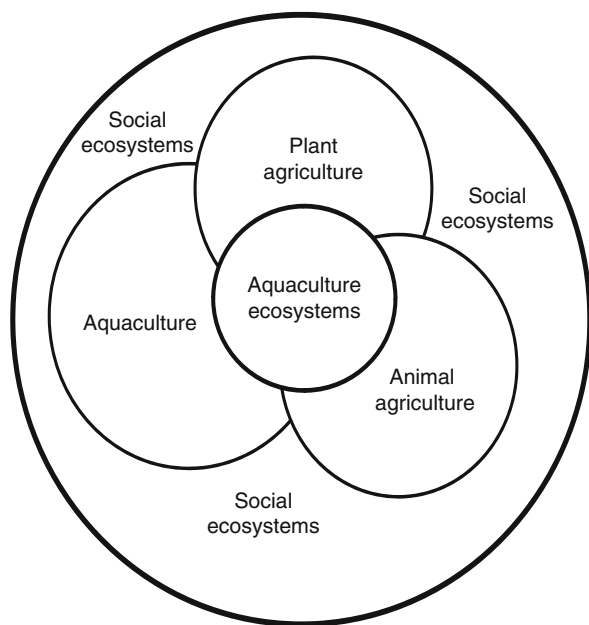
### Definition of Ecological Aquaculture

Ecological aquaculture is an alternative model of aquaculture development that uses ecological principles as the paradigm for the development of aquaculture [1, 2]. Ecological aquaculture plans, designs, develops, monitors, and evaluates aquatic farming ecosystems that preserve and enhance the form and functions of the natural and social environments in which they are situated. Ecological aquaculture farms are integrated “aquaculture ecosystems” designed to deliver both economic and social profit (Fig. 1).

Ecological aquaculture incorporates at the outset – and not as an afterthought – planning for not only the sustainable production of aquatic foods, but also for innovation [3], community development, and the wider social, economic, and environmental contexts of aquaculture at diverse scales, both large and small, and at the commercial, school, and homeowner scales [4, 5]. Ecological aquaculture also uses the “aquaculture toolbox” [6] to play vital roles in non-food, natural ecosystem rehabilitation, reclamation, and enhancement.

### Introduction

The roots of ecological aquaculture are in Asia [7, 8]. In this century, however, Asia, especially China, during the period from 1980s to present has chosen the industrial model of aquaculture development, and has dismantled much of its rich ecological aquaculture heritage, and choosing instead to intensify and import vast quantities of feedstuffs. As a result of intensification and the use of imported feeds, freshwater



**Aquaculture, Ecological. Figure 1**

Aquaculture ecosystems mimic the form and functions of natural ecosystems, but are a sophisticated, knowledge-based designed farming ecosystem that are planned as combinations of land and water-based aquatic plant, agronomic, algal, and animal subunits, which are embedded into the larger context of human social systems

aquaculture yields from China have increased 10X in just 20 years, and comprise the world's largest aquaculture industries [9].

The FAO ecosystems approach to aquaculture [10] creates a new code for global aquaculture development, combining into one common framework the two most important social-ecological trajectories for global aquaculture – aquaculture for the world's rich, and aquaculture for the world's poor. Knowledge of the rich archeology and anthropology of aquaculture connects this FAO code to antiquity, creating a single development pathway for aquaculture throughout human history.

### Key Principles

There are seven principles of ecological aquaculture:

1. **Ecological aquaculture systems are “aquaculture ecosystems” that mimic the form and functions of natural ecosystems.**

Ecological aquaculture farms are designed, farming ecosystems. Sophisticated site planning occurs so that farms “fit with nature” and do not displace or disrupt invaluable natural aquatic ecosystems or conservation areas. If localized displacement or degradation does occur, active support of innovative, collaborative research and development programs for ecosystems redesign, relocation, rehabilitation, and enhancement efforts are initiated and supported by the ecological aquaculture farms throughout the life of their farming operations.

2. **Ecological aquaculture is integrated with communities to maximize not only local but also regional economic and social multiplier effects in order to provide maximal job creation and training, and create “aquaculture communities” that are an essential part of vibrant, working waterfronts.**

Ecological aquaculture operations export to earn profits but also promote and market products locally to contribute to the development of society. Ecological aquaculture operations are committed to building the “culture” of aquaculture in order that “aquaculture communities” can develop and evolve as a source of innovation, education, and local pride. Aquaculture development as a means of community development can result in numerous, innovative economic and social multiplier effects such as aquaculture restaurants, marketing of “sustainable seafoods” that are branded as local and bioregional, and aquaculture tourism.

3. **Ecological aquaculture results in economic profit by practicing trophic efficiency to ensure that aquaculture is humanity's most efficient protein producer.**

Non-fed, shellfish and algae culture are preferred choices for ecological aquaculture developments. In fed aquaculture, fish meals/oils are not used as either the major protein or energy sources, but are included in animal diets to solve issues of diet palatability only; and, if used, fish meals and oils originate from certified, sustainable fishmeal/oil fisheries only. Fed aquaculture ecosystems rely on protein and oil sources from agricultural sources and seafood processing wastes, and include science innovations such as the development of detrital

food webs (“bioflocs”) to feed cultured, aquatic organisms.

4. **Ecological aquaculture results in social profit by integrating aquaculture developments into global fisheries, food, and poverty alleviation programs.**

Ecological aquaculture is part of the global movement to eliminate extreme hunger and starvation (Millennium Development Goal #1) by being a part of comprehensive plans for sustainable fisheries for poverty alleviation. Ecological aquaculture uses alternative feeds to support programs to deliver more of the world’s feed fisheries (sardines, anchovies, mackerels, etc.) away from aquaculture to the world’s poor.

5. **Ecological aquaculture practices nutrient management by using ecosystems design, reuse, and recycling, and does not discharge any nutrient or chemical pollution causing irreversible damage to natural aquatic or terrestrial ecosystems.**

No harmful metals, chemicals, or pharmaceuticals potentially harmful to long-term human or ecosystem health are used in the ecological aquaculture production processes. Ecological aquaculture farms have “sustainability strategic and implementation plans” in place to develop comprehensive, full cycle reuse, and recycling systems for all farming operations.

6. **Ecological aquaculture uses native species/strains, and does not contribute to “biological” pollution.**

Escapes from aquaculture, especially aquarium operations, have severely impacted aquatic ecosystems worldwide. Exotics species/strains can be good choices only if long-term monitoring data and scientific research indicate that exotic species are unlikely to establish; if exotic species are widely established and provide economic and social profit without irreversible environmental harm; or, the use of native species puts at risk indigenous genetic diversity. Ecological aquaculture operations ensure that innovative engineering and complete escapement technologies are used; that control and recovery procedures are in place; that active research and development programs provide alternatives and new options; and that complete, transparent, public documentation and information are available.

7. **Ecological aquaculture is a global partner, producing information for the world, avoiding the proprietary.**

Ecological aquaculture farms are aquaculture ecosystems that go beyond “meeting the regulations.” They are sites of collaboration, leadership development, and innovation. They are outstanding community citizens and models of stewardship [4]. Successful leadership development triggers developments of innovation and more efficient aquaculture-related technology, and more ecologically appropriate legislation and regulations.

**The FAO Ecological Approach to Aquaculture (EAA)**

In 2006, the Fisheries and Aquaculture Department of Food and Agriculture Organization (FAO) recognized the need to develop an ecosystem-based management approach to aquaculture similar to the Code of Conduct for Responsible Fisheries. FAO [10] suggested that an ecological approach to aquaculture (EAA) would have three main objectives: human well-being, ecological well-being, and the ability to achieve both via effective governance, within a hierarchical framework that was scalable at the farm, regional, and global levels.

In 2008, FAO defined an EAA as: A strategy for the integration of the activity within the wider ecosystem such that it promotes sustainable development, equity, and resilience of interlinked social-ecological systems. An ecosystem approach to aquaculture, similar to other systems approaches to management, accounts for a complete range of stakeholders, spheres of influences, and other interlinked processes. Applying an ecosystem-based approach must plan for physical, ecological, social, and economic systems as a part of community development, taking into account stakeholders in the wider social, economic, and environmental contexts of aquaculture [10]. FAO developed three principles and key issues at different scales of society:

**Principle 1: Aquaculture development and management should take account of the full range of ecosystem functions and services, and should not threaten the sustained delivery of these to society.**

The key issue is to estimate *resilience capacity*, or the limits to “acceptable environmental change.”



A range of terms has been used to estimate the limits to environmental change, including “environmental carrying capacity,” “environmental capacity,” “limits to ecosystem function,” “ecosystem health,” “ecosystem integrity,” “fully functioning ecosystems,” all of which are subject to a specific social/cultural/political context [11]. Conventional environmental impact assessments touch on just some of these issues. Application of the precautionary approach is important but inadequate in aquaculture; use of aquaculture risk assessment is becoming widespread [12].

**Principle 2: Aquaculture should improve human well-being and equity for all relevant stakeholders.**

Aquaculture should provide equal opportunities for development, which requires its benefits to be more widely shared especially locally so that it does not bring detriment to any sector of society, especially the poor. Aquaculture should promote both food security and safety as key components of human well-being.

**Principle 3: Aquaculture should be developed in the context of other sectors, policies, and goals.**

Interactions between aquaculture and its influences on the surrounding natural and social environment must be recognized. Aquaculture often has a smaller impact than other human activities, e.g., agriculture and industry, but it does not take place in isolation. There are many opportunities to couple aquaculture activities with other primary producing sectors in order to promote materials and energy recycling, and the better use of resources in general.

**Applying an Ecological Aquaculture Approach at Different Scales of Society**

There are three physical scales important in the planning for and assessment progress toward an ecosystem approach to aquaculture: farm scale, watershed/aquaculture zone, and global. Each of these has important planning and assessment needs.

**Farm Scale**

Planning for aquaculture farms is easily defined physically and could be few meters beyond the boundaries

of farming structures; however, the increasing size and intensity of some farms (e.g., large-scale shrimp farming or salmon farming) could affect an entire water body or watershed. Assessment of an EAA at the farm scale entails an evaluation of planning and implementation of “triple bottom line” programs – ecological, economic, and social programs – that in a comprehensive manner account for impacts to the wider ecosystem and social impacts of farm-level aquaculture developments, including use of better (“best”) management practices, and use of restoration, remediation, and mitigation methods. Proper site selection, levels of production intensity, use of species (exotic vs. native), use of appropriate farming systems technologies, and knowledge of economic and social impacts at the farm level should be considered.

For fed aquaculture, there are many concerns as the current trajectory and growth of the large-scale aquaculture industries. Current aquaculture development models are being modified rapidly by advances that will affect the widespread adoption of ecological aquaculture, which, if projected to 2050, confirm that large-scale aquaculture may move fully toward ecological aquaculture approaches (Table 1). There are a growing number of well-documented success stories in ecological aquaculture (Table 2).

**Watershed/Aquaculture Zone Scale**

Planning for an EAA at watersheds/aquaculture scale is relevant to common ecosystem and social issues such as diseases, trade in seed and feeds, climatic and landscape conditions, urban/rural development, etc. Assessment of an EAA at this scale is a two-phase process and will include, at phase I, assessments of

1. Inclusion of aquaculture as a part of regional governance frameworks, e.g., the overall framework of integrated coastal zone management or integrated watershed, land-water resource management planning, and implementation. Assessments take into account existing scenarios, user competition, and conflicts for land and water uses, and comparisons of alternatives for human development.
2. Impacts of aquaculture on regional issues such as escapees, disease transmission, and sources of contamination to/from aquaculture.

**Aquaculture, Ecological. Table 1** Major issues with fed aquaculture today (2010) and projections of these to 2050

| Issues                     | Concerns                                                                                                                                                                    | Modern developments (2010)                                                                                                                                                       | Trajectory of issues to 2050                                                                                                                                                            |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Feeds/no net gain          | Schroeder [28] documents pond can be a net consumer rather than a producer of animal protein. Fishing down and farming up marine food webs (Naylor et al. [29]; Pauly [30]) | Food conversion rates improve to ~1:1; fish in/out (FIFO) ratios drop to ~1.7; domestication of farmed species turns carnivores into domesticated omnivores                      | FIFO ratios drop to 1 or less; aquaculture uses ~50% of world's fish meal and oil with balance met by agricultural meals/oils                                                           |
| Feeds/ocean sustainability | Integrity of marine ecosystems damaged by high removal rates of feed species                                                                                                | Aquaculture use dropping due to rapid cost increases in meals/oils; poverty/social concerns recognized                                                                           | Ecosystem modeling parcels out science-based removal rates/allocations for aquaculture and ecosystems                                                                                   |
| Feeds/poverty              | Massive poverty and hunger in fish meal/oil producing countries                                                                                                             | New recognition in Peru; new international attention to role of meal/oil fisheries & fed aquaculture in poverty alleviation                                                      | Governments move to develop food products/prioritize human needs                                                                                                                        |
| Habitat destruction        | Mangrove destruction and water diversions disrupt nearshore and riverine ecosystems (Macintosh and Phillips [31]; Pullin et al. [32])                                       | Some nations (ex. Thailand) develop policies to prevent damage by proper siting and to rehabilitate damage of shrimp farms                                                       | Governments worldwide ban developments in sensitive conservation areas; widespread use of carrying capacity models (McKindsey et al. [33]) and ecological valuation for decision-making |
| Eutrophication             | Intensive aquaculture operations are feedlots producing nutrient pollution loads comparable to human sewage (Folke et al. [21]; Costa-Pierce [34])                          | Complete feeds, automated feed delivery systems, and nutrition research deliver less pollution; wastes are primarily in the form of soluble nutrients and feces, not waste feeds | Development of land-based recirculating systems; widespread use of land-based integrated aquaculture and water-based IMTA systems                                                       |
| Energy                     | Intensive aquaculture operations are energy intensive comparable to industrial agriculture and fisheries (Weatherly and Cogger [35])                                        | Scattered R&D in energy use, mostly Life Cycle Analyzes in aquaculture; little/no movement toward use of renewables                                                              | Renewable energy systems used                                                                                                                                                           |

3. Social considerations such as comprehensive planning for all of the possible beneficial multiplier effects of aquaculture on jobs and the regional economy, and considerations of aquaculture's impacts on indigenous communities.

At phase II, progress toward a full implementation of an EAA at watersheds/aquaculture zone scale can be assessed by measuring the

1. Abilities of governments to implement new methods of coastal and water governance to include ecological aquaculture
2. Development of ecological aquaculture approaches that allow agencies responsible for permitting aquaculture to consider and manage activities impacting aquaculture and aquatic ecosystems (e.g., capture fisheries, coastal zone development, watershed management organizations, agriculture, forestry, and industrial developments) more holistically, such as new mechanisms to communicate, cooperate, and collaborate across sectors
3. Design of ecological aquaculture management zones and parks that encourage aquaculture education, research, and development innovations and

**Aquaculture, Ecological. Table 2** Global success stories in ecological aquaculture

| Region/countries                                     | Aquaculture ecosystems                                                                                                                                                    | References                                                                    |
|------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Asia (China, Vietnam, Indonesia)                     | Rice-fish culture benefits millions of rural people; rice-fish aquaculture ecosystems have been designated as a “Globally important Agricultural Heritage System” (GIAHS) | FAO [9]; Dela Cruz et al. [36]                                                |
| Asia (China, Thailand, Cambodia, Vietnam, Indonesia) | Integrated aquaculture benefits millions of rural people                                                                                                                  | Edwards [8]                                                                   |
| Asia (China)                                         | Integrated Multi-trophic Aquaculture (IMTA) of fish, shellfish, and seaweeds bioremediates and increases total yields up to 50%                                           | Zhou et al. [37]                                                              |
| Egypt                                                | Integrated aquaculture produced over 650,000 tons of tilapia in 2008, ~60% of total fish production; provision of cheap source of fish at approx. same cost as poultry    | McGrath [38]                                                                  |
| Canada                                               | IMTA has been adopted by Cooke Aquaculture, the largest salmon aquaculture company in eastern Canada                                                                      | Chopin et al. [39]; Chopin [40]; Ridler et al. [41, 42]                       |
| Canada & USA                                         | Shellfish aquaculture has become widely accepted as environmentally friendly and socially acceptable                                                                      | National Academy [43, 44]                                                     |
| Tanzania                                             | Seaweed and shellfish aquaculture                                                                                                                                         | Seaweed grown by ~2,000 producers most women; new half-pearl industry growing |

partnerships, and also emphasize streamlined permitting of integrated aquaculture, polyculture, or innovative, integrated aquaculture–fisheries businesses and initiatives

### Global Scale

Planning for an EAA at a global scale considers aspects of transnational and multinational issues for global commodities (e.g., salmon and shrimp). Assessment of progress toward an EAA at the global level entails evaluation of issues such as: availabilities of fisheries and agriculture feedstocks for formulating aquaculture feeds and impacts on distant marine and social ecosystems, the economic and social impacts of aquaculture on fisheries and agriculture resources, impacts of aquaculture on markets, and impacts of globalization on social sustainability (social capital, goods, and social opportunities). Applications of tools such as lifecycle assessments of aquaculture commodities and

the use of innovative social enterprise management guidelines and tools are useful to determine impacts at the global scale.

### Social Ecology of Aquaculture

Many analysts are calling for more integrated, multidisciplinary ways of developing ecologically and socially responsible food, energy, water, and waste systems to meet society’s needs [13]. Among the first was Lubchenco [14] who called for a new social contract for science and society. Industrial aquaculture in its current development phase does not have a social contract or social license to expand in many areas of the world, especially at the watershed/aquaculture zone and global scales.

Just as important are social investments in aquaculture at the individual level. Aquaculture has an urgent need for developing and engaging leaders who are well trained and experienced decision-makers who are

“honest brokers of policy alternatives” [15]. Keen et al. [16] believe transformation toward more sustainable practices will be much more likely if the individuals who make up society can accept change and modify their personal behaviors [17]. Changes in the behavior of individuals can “scope up” and result in larger changes at the community and societal scales by employing a combination of trust-building, favorable performance, accountability, flexibility and innovation, and the inclusion of stakeholders in strategic planning [18, 19].

Folke et al. [20] challenge our education system to continually adapt to the emergence of such new questions and changing social compacts as aquaculture. Any rapid progress toward an ecological approach to aquaculture will require development of education programs that promote broad awareness, recognition, and implications of new approaches to aquaculture, and the creation of new institutions. Bransford et al. [22] suggest that for such subfields of sustainability science as aquaculture more attention needs to be given to educating the next generation of leaders by teaching metacognitive skills such as practicing different ways of thinking in a variety of contexts, with less emphasis being placed on trying to fill students with a large volume of facts and knowledge.

### **An Ecological Aquaculture Strategy for the “Triple Bottom Line”**

Aquaculture development plans will be incomplete unless both economic and social goals are articulated, and agreed upon, at the outset, in transparent, participatory processes. Only then can aquaculture “evolve” as an integral part of – not separate from – farmers, fishermen, sustainable community development, and the future of “working waterfronts.” Aquaculture’s success cannot simply be defined as having successfully developed the hatchery, feed, and marketing components of a business plan – the old alignment of the “seed, feed, and the need.” Rather, sustainable, ecological aquaculture nurtures “society’s success” for the “triple bottom line” of economic, environmental, and social profit [23].

Adversarial social processes occur in jurisdictions where aquaculture is not being developed using

a social-ecological “ecosystem approach.” In these places, the blue revolution is being televised, tweeted, and blogged. Adversarial processes (conflicts) occur when stakeholders do not recognize each other’s interests as legitimate. These processes increase conflict; thrive on uncertainty; have poor communication; are exclusive, divisive, opaque, and closed, and lack trust. Collaborative processes must be created that create trust through shared learning and ownership, creative problem solving, joint fact finding, and employ adaptive management. Robertson and Hull [24] call this a “public ecology” that has both process and content that emphasizes the participation of extended peer communities of research specialists, policy-makers, and concerned citizens. Dasgupta and Maler [25] have used tools developed by economists and ecologists to value choices in the midst of this complexity. In general, since aquaculture is such a dynamic, evolutionary field, managers, policy-makers, and community leaders need to participate to allow understanding of new and emerging problems and to stimulate multidisciplinary research; as analysts report that such work is the highest impact science being published today [26].

Clear, unambiguous linkages between aquaculture and the environment must be created and fostered, and the complementary roles of aquaculture in contributing to environmental sustainability, rehabilitation, and enhancement must be developed and clearly articulated to a highly concerned, increasingly educated, and involved public. New aquaculture operations must plan, at the outset, to:

1. Become an integral part of a community and a region.
2. Plan for community development by working with leaders to provide needed inputs and recycle wastes.
3. Create a diversity of unprocessed and value-added products, and provide local market access, since in rich societies, aquaculture products are high-value discretionary purchases that can easily be rejected by the public.
4. Plan for job creation and environmental enhancement on both local and regional scales.

It is well documented that most aquaculture jobs are not directly in production, rather in the affiliated service industries. In the USA, Dicks et al. [27] found

that aquaculture production accounted for just 8% of the income and ~16,500 jobs. Aquaculture goods and services accounted for 92% of the income and ~165,500 jobs (most jobs were in equipment, supplies, feeds, fertilizers, transport, storage, processing). However, most aquaculture development plans focus almost exclusively on production concerns and have little/no comprehensive plans for localization of seed, feed, markets, or other aquaculture service industries that produce the most benefits to local economies – to say nothing about employing local professionals (most industrial aquaculture operations import high paying professionals from the outside). In the vast majority of cases, feed and services are imported to sites, and local people cannot even buy the produce!

An ecological aquaculture development model will create new opportunities for a wider group of professionals to get involved in aquaculture since new advances will be needed not only in technology but also in information, community development, and facilitation. Ecological aquaculture as a “new” field, and one that is important for the future food security and environment of the planet, requires the much more comprehensive planning in order to evolve a new social contract with society.

### Future Directions

By 2030, fed aquaculture will turn from the ocean to land-based agriculture to provide its feeds and oils. As such, more sophisticated, ecologically designed and integrated aquaculture systems will become more widespread because they better fit the social-ecological context of both rich and poor countries. Ecological aquaculture provides the basis for developing a new social contract for aquaculture that is inclusive of all stakeholders and decision-makers in fisheries, agriculture, ecosystems conservation, and restoration.

The wildly optimistic scenarios for aquaculture’s expansion will not occur unless alternative ecological approaches and ecological intensification of aquaculture are widely adopted. Aquaculture needs to be better integrated into overall fishery societal plans for securing sustainable seafood supplies and restoring damaged, supporting fisheries ecosystems.

The overuse and degraded state of nearly all of the world’s aquatic ecosystems, combined with public

concerns about adding any “new” uses or sources of aquatic pollution to already overburdened natural and human systems, require aquaculture to develop ecosystems approaches; sustainable operating procedures; and to articulate a sustainable, ecological pedagogy. The fact that an ecological aquaculture approach can ensure aquaculture is a net gain to humanity; and it could be the key organizing paradigm to form a new social contract for aquaculture worldwide.

The massive globalization of seafood trade has meant less dependence on local natural and social ecosystems, and has resulted in some virulent opposition to aquaculture development, especially as industrial aquaculture has removed the local sources of production and markets, and jobs have been externalized. One major consequence of this globalization has been the increased dependence of industrial, “fed” aquaculture on the southeastern Pacific Ocean marine ecosystem for fish meals and oils. The global implications for the Humboldt ecosystem, for local poverty, and the scoping of this unsustainable situation to the entire global protein food infrastructure are profound, and are still largely unrealized.

Aquaculture sites are not only economic engines of primary production that meet the regulations of a society, but can be sites of innovation and pride if they can be well designed as community-based, aquaculture farming ecosystems. A review of the progress toward such an ecosystems approach to aquaculture is necessary to inspire planners and environmental decision-makers at many societal scales (national, regional, local) to make use of such innovative approaches. Sophisticated site planning of aquaculture can occur so that farms “fit with nature” and do not displace or disrupt invaluable natural, aquatic ecosystems, or conservation areas; but contribute to the local economy and society [5].

An ecological aquaculture approach to comprehensive planning for aquaculture at many different scales will integrate aquaculture into plans for not only environmental benefit and the restoration of coastal ecosystems, but also local market developments and the future of coastal communities. As such, ecological aquaculture can move aquaculture beyond endless user conflicts, and could stabilize the regulatory environment and ensure a more equitable process of eco-social design of aquaculture for the future.



## Bibliography

### Primary Literature

- Costa-Pierce B (2002) Ecological aquaculture. Blackwell, Oxford
- Costa-Pierce B (2003) Use of ecosystems science in ecological aquaculture. *Bull Aquacul Assoc Canada* 103(2):32–40
- Culver K, Castle D (2008) Aquaculture, innovation and social transformation. Springer, New York
- Costa-Pierce B (2008a) An ecosystem approach to marine aquaculture: a global review. In: Soto D (ed) Building an ecosystem approach to aquaculture, vol 14, FAO fisheries and aquaculture proceedings. FAO, Rome, pp 81–116
- Costa-Pierce B (2008b) Epilogue. In: Culver K, Castle D (eds) Aquaculture, innovation and social transformation. Springer, New York, pp 315–325
- Costa-Pierce B, Bridger C (2002) The role of marine aquaculture facilities as habitats and ecosystems. In: Stickney R, McVey J (eds) Responsible marine aquaculture. CABI, Wallingford, pp 105–144
- Ruddle K, Zhong G (1988) Integrated agriculture-aquaculture in South China: the Dike-Pond system in Zhujiang Delta. Cambridge University Press, Cambridge
- Edwards P (2009) Traditional Asian aquaculture: definition, status and trends. In: Burnell G, Allan G (eds) New technologies in aquaculture, improving production efficiency, quality and environmental management. Woodhead Publishing Ltd., Cambridge, UK, pp 1029–1063
- FAO (Food and Agriculture Organization) (2009) The state of world fisheries and aquaculture 2008. Fisheries Department, Food and Agriculture Organization of the United Nations, Rome
- Soto D et al (2008) Applying an ecosystem-based approach to aquaculture: principles, scales and some management measures. In: Soto D, Aguilar-Manjarrez J, Hishamunda N (eds) Building an ecosystem approach to aquaculture, vol 14, Rome: FAO fisheries and aquaculture proceedings. FAO, Rome
- Hambrey J, Senior B (2007) Taking forward environmental carrying capacity and ecosystem services. Recommendations for CCW. CCW Policy Research Report No. 07/22
- Black E (2006) A risk analysis approach to assessment of the potential genetic interactions of non-salmonid marine finfish escapes from aquaculture with local native wild stocks, pp 93–105. In: ICES WGEIM report on the environmental impacts of mariculture, 24–28 April 2006, Narragansett, Rhode Island, USA. ICES, Copenhagen, Denmark. (Available at [www.ices.dk/reports/MCC/2006/WGEIM06.pdf](http://www.ices.dk/reports/MCC/2006/WGEIM06.pdf))
- Dees J, Backman E (1994) Social enterprise: private initiatives for the common good. Harvard Business School Publishing, Harvard University, Boston
- Brown L (2009) Plan B 4.0: mobilizing to save civilization. W. Norton, New York
- Lubchenco J (1998) Entering the century of the environment: a new social contract for science. *Science* 279:491–497
- Pilke R (2007) The honest broker: making sense of science in policy and politics. Cambridge University Press, Cambridge
- Keen M, Brown V, Dyball B (2005) Social learning in environmental management. Earthscan, London
- Huckle J, Sterling S (1996) Education for sustainability. Earthscan, London
- Brehm J, Rahn W (1997) Individual level evidence for the causes and consequences of social capital. *Am J Pol Soc* 41:999–1023
- Knack S (2002) Social capital and the quality of government: evidence for the states. *Am J Pol Sci* 46:772–785
- Folke C, Hahn T, Olsson P, Norberg J (2005) Adaptive governance of social-ecological systems. *Ann Rev Environ Res* 30:441–473
- Bransford J, Brown A, Cocking R (2000) How people learn: brain, mind, experience, and school. National Academy Press, Washington
- Savitz A (2006) The triple bottom line: how today's best-run companies are achieving economic, social and environmental success – and how you can too. Jossey-Bass/Wiley, New York
- Robertson D, Hull R (2003) Public ecology: an environmental science and policy for global society. *Environ Sci Policy* 6:399–410
- Dasgupta P, Maler K (2004) The economics of non-convex ecosystems. Springer, New York
- Jones B, Wuchty S, Uzzi B (2008) Multi-university research teams: shifting impact, geography, and stratification in science. *Science* 322:1259–1262
- Dicks M, McHugh R, Webb B (1996) Economy-wide impacts of U.S. aquaculture. Oklahoma Agricultural Experiment Station, Norman, 946p
- Schroeder G (1980) Fish farming in manure loaded ponds. In: Pullin R, Shehadeh Z (eds) Integrated agriculture-aquaculture farming systems. ICLARM, Manila, pp 73–86
- Naylor R, Goldburg R, Primavera J, Kautsky N, Beveridge M, Clay J, Folke C, Lubchenco J, Mooney H, Troell M (2000) Effect of aquaculture on world fish supplies. *Nature* 405:1017–1024
- Pauly D (2000) Global change, fisheries, and the integrity of marine ecosystems: the future has already begun. In: Pimental D, Westra L, Noss R (eds) Ecological integrity. Island Press, Washington, DC, pp 227–244
- Macintosh D, Phillips M (1992) Environmental considerations in shrimp farming. *Infotish Int* 6:38–92
- Pullin R (1993) Environment and aquaculture in developing countries. International Center for Living Aquatic Resources Management, Manila
- McKinsey C, Thetmeyer H, Landry T, Silvert W (2006) Review of recent carrying capacity models for bivalve culture and recommendations for research and management. *Aquaculture* 261(2):451–462
- Costa-Pierce B (1997) Environmental impacts of nutrients discharged from aquaculture: towards the evolution of sustainable, ecological aquaculture systems. In: Baird D, Beveridge M, Kelly L, Muir J (eds) Aquaculture and water resource management. Blackwell, Oxford, UK, pp 81–113

35. Weatherly A, Cogger B (1977) Fish culture: problems and prospects. *Science* 197:427–430
36. Dela Cruz C, Lightfoot C, Costa-Pierce B, Carangal V, Bimbao M (1992) Rice-fish research and development in Asia. ICLARM, Manila
37. Zhou Y, Yang H, Hu H, Liu Y, Mao Y, Zhou H, Xu X, Zhang F (2006) Bioremediation potential of the macroalga *Gracilaria lemaneiformis* (Rhodophyta) integrated into fed fish culture in coastal waters of north China. *Aquaculture* 252:264–276
38. McGrath C (2009) Egypt. A big catch feeds millions. *Global geopolitics Net Sites/IPS* <http://globalgeopolitics.net/wordpress/2009/11/01/egypt-a-big-catch-feeds-millions>
39. Chopin T, Buschmann A, Halling C, Troell M, Kautsky N, Neori A, Kraemer G, Zertuche-Gonzalez J, Yarish C, Neefus C (2001) Integrating seaweeds into marine aquaculture systems: a key toward sustainability. *J Phycol* 37:975–986
40. Chopin T (2006) Integrated multi-trophic aquaculture. What it is, and why you should care. . . and don't confuse it with polyculture. *Northern Aquacult* 12(4):4
41. Ridler N, Robinson B, Chopin T, Robinson S, Page F (2006) Development of integrated multi-trophic aquaculture in the Bay of Fundy, Canada: a socio-economic case study. *World Aquacult* 37(3):43–48
42. Ridler N, Wowchuk M, Robinson B, Barrington K, Chopin T, Robinson S, Page F, Reid G, Haya K (2007) Integrated multi-trophic aquaculture (IMTA): a potential strategic choice for farmers. *Aquacul Econ Manag* 11:99–110
43. National Research Council (2009) Shellfish mariculture in Drakes Estero, Point Reyes national seashore. National Academies Press, Washington, California
44. National Research Council (2010) Ecosystem concepts for sustainable bivalve mariculture. National Academies Press, Research Council, Washington

## Books and Reviews

- Alder J, Campbell B, Karpouzi V, Kaschner K, Pauly D (2008) Forage fish: from ecosystems to markets. *Ann Rev Envir Res* 33:71–714
- Bell J, Waagbo R (2002) Safe and nutritious aquaculture produce: benefits and risks of alternative sustainable aquafeeds. In: Holmer M, Black K, Duarte C, Marba N, Karakassis I (eds) *Aquaculture in the ecosystem*. Springer, New York, pp 185–226
- Black K (2001) Sustainability of aquaculture. In: Black K (ed) *Environmental impacts of aquaculture*. Sheffield Academic Press, Sheffield, pp 199–212
- Costa-Pierce B, Desbonnet A, Edwards P, Baker D (eds) (2005) *Urban Aquaculture*. CABI Publishing, Wallingford, UK
- Duarte C, Holmer M, Olsen Y, Stot D, Maba N, Guiu J, Black K, Karakassis I (2009) Will the oceans help feed humanity? *Bioscience* 59:967–976
- Gowen R, Bradbury N (1987) The ecological impact of salmonid farming in coastal waters: a review. *Ocean Mar Biol Ann Rev* 25:563–575
- Stickney R, McVey J (2002) *Responsible marine aquaculture*. CABI, Wallingford

- Tacon A, Hasan M, Subasinghe R (2006) Use of fishery resources as feed inputs for aquaculture development: trends and policy implications. *FAO fisheries circular No. 1018*. Food and Agriculture Organization of the United Nations, Rome
- Tacon A, Metian M (2008) Global overview on the use of fish meal and fish oil in industrially compounded aquafeeds: trends and future prospects. *Aquaculture* 285:146–158
- Turchini G, Bente E, Torstensen E, Ng W (2009) Fish oil replacement in finfish nutrition. *Rev Aquacult* 1:10–57

---

## Aquaculture, Integrated Multi-trophic (IMTA)

THIERRY CHOPIN

Canadian Integrated Multi-Trophic Aquaculture Network (CIMTAN), University of New Brunswick, Saint John, NB, Canada

## Article Outline

Glossary

Definition of the Subject

Introduction: Aquaculture Is Needed But Some Practices Need to Evolve

IMTA: A Flexible and Functional Concept

The Need for Diversifying Responsible Aquaculture Systems and for an Ecosystem Approach

IMTA, While Not the Panacea to and for Everything, Is, Nevertheless, One of the Improvement Options

Recognizing and Valuing the Biomitigative Services Rendered by the Extractive Components of IMTA: Should a System of Nutrient and Carbon Trading Credits Be Developed?

What Will It Take to Increase the Acceptance and Adoption of IMTA as a Responsible Aquaculture Practice of the Future?

Future Directions: The Path Forward

Acknowledgments

Bibliography

## Glossary

**Biomitigative services provided by extractive aquaculture** The environmental, economic, and societal services and benefits received by ecosystems – in their broad definition which includes humans who

depend on them – from the conditions and processes of cultivated species, such as seaweeds extracting inorganic nutrients and suspension- and deposit-feeders extracting organic particles recaptured from the activities of fed aquaculture (e.g., fish or shrimp aquaculture), to maintain their health. Biomitigative services can also be provided by natural populations of similar organisms.

**Integrated multi-trophic aquaculture (IMTA)** The farming, in proximity, of aquaculture species from different trophic levels, and with complementary ecosystem functions, in a way that allows one species' uneaten feed and wastes, nutrients, and by-products to be recaptured and converted into fertilizer, feed, and energy for the other crops, and to take advantage of synergistic interactions between species. Farmers combine fed aquaculture (e.g., finfish or shrimps) with extractive aquaculture, which utilizes the inorganic (e.g., seaweeds or other aquatic vegetation) and organic (e.g., suspension- and deposit-feeders) excess nutrients from fed aquaculture for their growth. The aim is to ecologically engineer balanced systems for environmental sustainability (biomitigative services for improved ecosystem health), economic stability (improved output, lower costs, product diversification, risk reduction, and job creation in disadvantaged communities) and societal acceptability (better management practices, improved regulatory governance, and appreciation of differentiated and safe products).

### Definition of the Subject

Fulfilling aquaculture's growth potential requires responsible technologies and practices. Sustainable aquaculture should be ecologically efficient, environmentally benign, product-diversified, profitable, and societally beneficial. Integrated multi-trophic aquaculture (IMTA) has the potential to achieve these objectives by cultivating fed species (e.g., finfish or shrimps fed sustainable commercial diets) with extractive species, which utilize the inorganic (e.g., seaweeds or other aquatic vegetation) and organic (e.g., suspension- and deposit-feeders) excess nutrients from fed aquaculture for their growth. Thus, extractive aquaculture produces valuable biomass, while simultaneously rendering

biomitigative services for the surrounding ecosystem and humans. Through IMTA, some of the uneaten feed and wastes, nutrients, and by-products, considered "lost" from the fed component, are recaptured and converted into harvestable and healthy seafood of commercial value, while biomitigation takes place (partial removal of nutrients and CO<sub>2</sub>, and supplying of oxygen). In this way, some of the externalities of fed monoculture are internalized, hence increasing the overall sustainability, profitability, and resilience of aquaculture farms. A major rethinking is needed regarding the definition of an "aquaculture farm" (reinterpreting the notion of site-lease areas) and regarding how it works within an ecosystem, in the context of a broader framework of Integrated Coastal Zone Management (ICZM). The economic values of the environmental/societal services of extractive species should be recognized and accounted for in the evaluation of the true value of these IMTA components. This would create economic incentives to encourage aquaculturists to further develop and implement IMTA. Seaweeds and invertebrates produced in IMTA systems should be considered as candidates for nutrient/carbon trading credits (NTC and CTC) within the broader context of ecosystem goods and services. Long-term planning/zoning promoting biomitigative solutions, such as IMTA, should become an integral part of coastal regulatory and management frameworks.

### Introduction: Aquaculture Is Needed But Some Practices Need to Evolve

The global seafood industry is at a crossroads: as capture fisheries stagnate in volume, they are falling increasingly short of a growing world demand for seafood. It is anticipated that by 2030, there will be a 50–80 million ton seafood deficit [1]. This gap will likely not be filled by capture fisheries but by aquaculture operations, which already supply almost 50% of the seafood consumed worldwide [1]. Consequently, it is imperative to design the ecosystem responsible aquaculture practices of tomorrow that maintain the integrity of ecosystems while ensuring the viability of this sector and its key role in food provision, safety, and security.

Without a clear recognition of the industry's large-scale dependency and impact on natural ecosystems and traditional societies, the aquaculture industry is

unlikely to either develop to its full potential, continue to supplement ocean fisheries, or obtain societal acceptance. The majority of aquaculture production still originates from relatively sustainable extensive and semi-intensive systems [2]; however, the rapid development, throughout the world, of intensive marine fed aquaculture (e.g., carnivorous finfish and shrimp) is associated with concerns about the environmental, economic, and social impacts that these, often mono-specific, practices can have, especially where activities are highly geographically concentrated or located in suboptimal sites whose assimilative capacity is poorly understood and, consequently, prone to being exceeded. There are also some concerns with shellfish aquaculture, especially at high density, as shellfish occupy an intermediate trophic level and often play a dual role of organic filtering organisms and waste/nutrient generating organisms [3].

For many marine aquaculture operations, monoculture is, spatially and managerially, often the norm. Species are cultivated independently in different bays or regions. Consequently, the two different types of aquaculture (fed versus extractive) are often geographically separate, rarely balancing each other out at the local or regional scale, and, thus, any potential synergy between the two is lost. To avoid pronounced shifts in coastal processes, the solution to nutrification by fed aquaculture is not dilution, but extraction and conversion of the excess nutrients and energy into other commercial crops produced by extractive aquaculture.

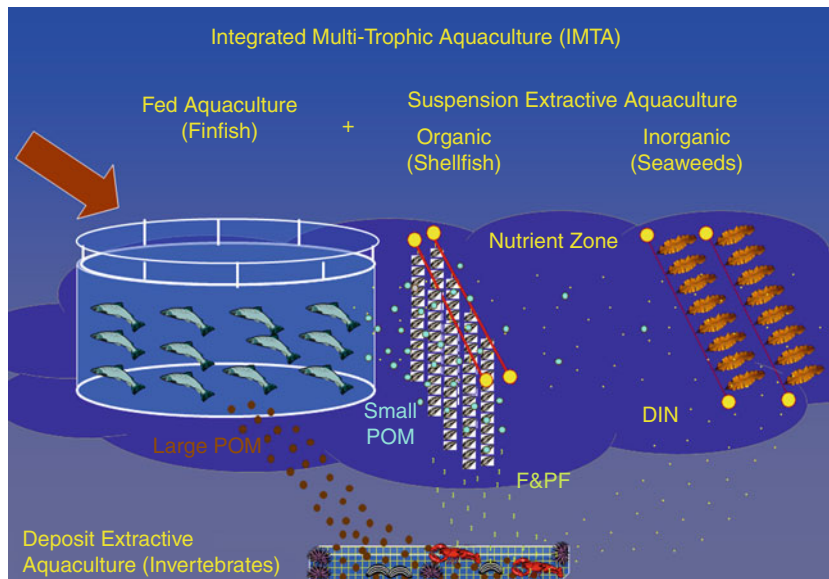
To continue to grow, while developing better management practices, the aquaculture sector needs to develop more innovative, responsible, sustainable, and profitable technologies and practices, which should be ecologically efficient, environmentally benign, product-diversified, and societally beneficial. Maintaining sustainability, not only from an environmental, but also from economic, social, and technical perspectives, has become a key issue, increased by the enhanced awareness of more and more demanding consumers regarding quality, traceability, and production conditions. Integrated multi-trophic aquaculture (IMTA) has the potential to play a role in reaching these objectives by cultivating fed species (e.g., finfish or shrimps fed sustainable commercial diets) with extractive species, which utilize the inorganic (e.g., seaweeds or other aquatic vegetation) and organic

(e.g., suspension- and deposit-feeders) excess nutrients from aquaculture for their growth (Fig. 1).

### IMTA: A Flexible and Functional Concept

The IMTA concept is extremely flexible [4]. To use a musicology analogy, IMTA is the central/overarching theme on which many variations can be developed according to the environmental, biological, physical, chemical, societal, and economic conditions prevailing in parts of the world where the IMTA systems are operating. It can be applied to open-water or land-based systems, marine or freshwater systems (sometimes called “aquaponics”), and temperate or tropical systems. What is important is that the appropriate organisms are chosen at multiple trophic levels based on the complementary functions they have in the ecosystem, as well as for their economic value or potential. In fact, IMTA is doing nothing other than recreating a simplified, cultivated ecosystem in balance with its surrounding instead of introducing a biomass of a single type one thinks can be cultivated in isolation from everything else. Integration should be understood as cultivation in proximity, not considering absolute distances but connectivity in terms of ecosystemic functionalities.

It should be made clear that in the minds of those who created the acronym “IMTA,” it was never conceived to be viewed with the minimalist perspective of only the cultivation of salmon (*Salmo salar*), kelps (*Saccharina latissima* and *Alaria esculenta*), and blue mussels (*Mytilus edulis*) within a few hundred meters: this is only one of the variations (Fig. 2) and the IMTA concept can be extended to very large ecosystems like the Yellow Sea (see below). This also means that IMTA variations include integrated agriculture aquaculture systems (IAAS), integrated silviculture (mangrove) aquaculture systems (ISiAS), integrated green water aquaculture systems (IGWAS), integrated peri-urban aquaculture systems (IPUAS), integrated fisheries aquaculture systems (IFAS), sustainable ecological aquaculture systems (SEAS), integrated temporal aquaculture systems (ITAS), and integrated sequential aquaculture systems (ISAS, also called partitioned aquaculture systems, PAS, or fractionated aquaculture systems, FAS) [5–7]. There is no ultimate IMTA system to “feed the world.” There is not one



**Aquaculture, Integrated Multi-trophic (IMTA). Figure 1**

Conceptual diagram of an Integrated Multi-Trophic Aquaculture (IMTA) operation including the combination of fed aquaculture (e.g., finfish) with suspension organic extractive aquaculture (e.g., shellfish), taking advantage of the enrichment in small particulate organic matter (POM); inorganic extractive aquaculture (e.g., seaweeds), taking advantage of the enrichment in dissolved inorganic nutrients (DIN); and deposit organic extractive aquaculture (e.g., echinoids, holothuroids, and polychaetes), taking advantage of the enrichment in large particulate organic matter (POM) and feces and pseudo-feces (F&PF) from suspension-feeding organisms. The bioturbation on the bottom also regenerates some DIN, which becomes available to the seaweeds



**Aquaculture, Integrated Multi-trophic (IMTA). Figure 2**

One of the Integrated Multi-Trophic Aquaculture (IMTA) sites in the Bay of Fundy, New Brunswick, Canada, operated by Cooke Aquaculture Inc.: two rows of salmon cages in the background, then a row of mussel rafts and two seaweed rafts in the foreground



world but climatic, environmental, biological, physical, chemical, economic, societal, and political conditions, each of which can lead to different choices of systems for feeding these microworlds.

The paradox is that IMTA is not a new concept. Asian countries, which provide more than two thirds of the world's aquaculture production, have been practicing IMTA (often described as a type of "polyculture") for centuries, through trial and error and experimentation. Why, then, is this common sense solution not more widely implemented? The reasons for this generally center around social customs and practices, and market-driven economic models not considering externalities that one is already familiar with, even if common sense tells one that one should modify them. Human society does not change quickly unless there are compelling reasons to do so. What to do when early large profit margins create short-term economic booms, followed within a few decades by dwindling meager profit margins? Often, the temptation is to throw more large volume cultivation operations and destructive methods into the mix, without proper regulations and business plans. Pollution, disease and economic busts generally ensue, major restructuring of the industry becomes necessary, and a few clairvoyant visionaries remain afloat and adapt to jump to the next curve to survive. This evolution is not exclusive to the aquaculture industry. Why do humans have such short and selective memories resulting in them repeating mistakes, regularly?

The fact that humans are currently at a crossroad should motivate them to improve current aquaculture practices, without further delay. Fishery management plans in most countries have been single-species approaches, completely neglecting the interactions between species, not understanding the synergies, or antagonisms, between them and how an ecosystem works based on the complementarities of the different functions of the different organisms inhabiting it. It seems that, despite the knowledge of the limitations of mono-agriculture and mono-fisheries, people are ready to develop similar plans for the management of mono-aquaculture. It should be recognized that there is still a chance for incorporating all the learning about the problems of terrestrial monocultures into the relatively new frontier of aquaculture. To better manage marine, brackish, or freshwater environments to the

benefits of mankind and the ecosystem, one needs to develop a new science, marine agronomy, learning from the mistakes made in land agriculture over the centuries to do a better job with aquaculture. It is interesting to note that traditional agricultural practices, such as crop alternation and fallow, are now being transposed to aquaculture practices.

Why, then, is IMTA not more widely adopted, especially in the western world? Paul Greenberg, in his fascinating book "Four Fish" [8], mentioned a very interesting point. In Leviticus, the third book of the Hebrew bible in which, according to the Jewish tradition, God dictated commandments to Moses, one can read (19:19): "You must not sow your field with two different kinds of seed" (also translated as "two kinds of seed" or as "mixed seed"). One can wonder if this represents, in fact, one of the most ancient treatises recommending mono-agricultural practices and if it is not the reason why integrated culture techniques have been ignored for centuries in the Judeo-Christian civilization, while they have flourished in other civilizations, especially in Asia. Moreover, if Asian cultures are accustomed to the concept of considering wastes from farming practices as resources for other crops rather than pollutants, this attitude still has a long way to progress in the western world where aquaculture is a more recent development.

### **The Need for Diversifying Responsible Aquaculture Systems and for an Ecosystem Approach**

The common old saying "Do not put all your eggs in one basket," which applies to agriculture and many other businesses, should also apply to aquaculture. Having excess production of a single species leaves a business vulnerable to sustainability issues because of fluctuating prices in what has become commodity markets and potential oversupply, and the possibility of catastrophic destruction of one's only crop (diseases, damaging weather conditions). Consequently, diversification of the aquaculture industry is advisable for reducing economic risk and maintaining sustainability and competitiveness.

From an ecological point of view, diversification also means cultivating more than one trophic level, i.e., not just raising several species of finfish (that

would be “polyculture”), but adding into the mix organisms of different and lower trophic levels (e.g., seaweeds, shellfish, crustaceans, echinoderms, worms, bacteria, etc.) to mimic the functioning of natural ecosystems. Staying at the same ecological trophic level will not address some of the environmental issues because the system will remain unbalanced due to nondiversified input and output needs. Evolving aquaculture practices will require a conceptual shift toward understanding the working of food production systems rather than focusing on technological solutions.

One of the innovative solutions promoted for environmental sustainability (biomitigative services for improved ecosystem health), economic stability (improved output, lower costs, product diversification, risk reduction, and job creation in disadvantaged communities), and societal acceptability (better management practices, improved regulatory governance, and appreciation of differentiated and safe products) is IMTA. The aim is to increase long-term sustainability and profitability per cultivation unit (not per species in isolation as is done in monoculture), as some of the uneaten feed and wastes, nutrients, and by-products of one crop (fed animals) are not lost but recaptured and converted into fertilizer, feed, and energy for the other crops (extractive plants and animals). These, in turn, can be harvested and marketed as healthy seafood, while feed costs are reduced because of their reuse in multiple niches and biomitigation is taking place (partial removal of nutrients and CO<sub>2</sub>, and supply of oxygen). In this way, all the cultivation components have a commercial value, as well as key roles in recycling processes and rendering biomitigative services. Some of the externalities of fed monoculture are internalized, hence increasing the overall sustainability, long-term profitability, and resilience of aquaculture farms. The harvesting of the different types of crops participates in the capture and export of nutrients outside of the coastal ecosystem.

The biomass and functions of the fed and extractive species naturally present in the ecosystem in which aquaculture farms are operating must also be accounted for or this will lead to the development of erroneous carrying/assimilative capacity models. For example, the 158,811 t (fresh weight) of the intertidal seaweed, *Ascophyllum nodosum* (rockweed), in proximity to salmon aquaculture operations in southwest

New Brunswick, Canada, are not neutral in the ecosystem and represent a significant coastal nutrient scrubber which should be taken into consideration to understand the functioning of that part of the Bay of Fundy.

### **IMTA, While Not the Panacea to and for Everything, Is, Nevertheless, One of the Improvement Options**

IMTA has never been portrayed as the solution to and for everything! For example, IMTA does not address the issues of escapees from open-water fish farms. It is, of course, in the interest of everybody, especially the industry (to not lose money) to reduce the number of escapees. This is, however, a question of engineering of the rearing systems (cages, netting material, etc.) and the suitability of the environment to survival should escapees occur. To solve the escapee issue, it has been suggested that fish farms should be pulled from the open water and placed on land or in closed containment. Moving on land is, however, not a guarantee for zero escapees. There are well-known escapee cases from land-based operations, with serious consequences. For example, the bighead carp (*Hypophthalmichthys nobilis*) and the silver carp (*Hypophthalmichthys molitrix*) were brought from Asia to the southern USA in the 1970s to help control algal proliferation in channel catfish (*Ictalurus punctatus*) farms. There are reports of escapees into the lower Mississippi River system, especially associated with flood episodes in the early 1990s. Self-sustaining populations have been able to move northward to enter the Upper Mississippi River system and the Illinois River system. Presently, there are fears that these fish could enter the Great Lakes system through the Chicago Sanitary and Ship Canal and the Des Plaines River to finally reach Lake Michigan, after an escape of around 2,000 km in approximately 20–30 years. Electric fish barriers have been put in place, but their efficiency has been questioned. The use of rotenone, a biodegradable piscicide, was authorized but seemed to have killed more common carps (*Cyprinus carpio*; itself an introduced species from Europe in the 1830s) than bighead and silver carps. On April 26, 2010, the US Supreme Court decided not to get involved in a dispute over how to prevent these carps from making their way into the Great Lakes; it turned down a new

request by the State of Michigan to consider ordering permanent closing of the Chicago-area shipping locks. What the impacts on the ecosystems could be, should these fish get into the Great Lakes systems, is unknown, but they are well-known for their ability to consume large amounts of algae and zooplankton, eating as much as 40% of their body weight per day, and they are fierce competitors when it comes to securing their food needs. The silver carp is also a danger to recreational fishers, water skiers, and boaters because of its habit to jump out of the water when startled by boat motors or other noises, creating life-threatening aerial hazards with high speed impacts.

The number of escapees from land-based facilities is not as well documented as with cage-based aquaculture. Perhaps because land-based fish escapes are more likely to occur as a continuous “trickle” instead of a single major event such as a net tear that would lead to “large-scale” escapes. However, reports do surface from time to time in the media, particularly if there is some novelty in the story. A recent example is the report of the cultured salmonid brown trout, *Salmo trutta*, escaping from a pond farm in the UK. A wildlife photographer caught them in action, making large leaps out of the water straight into a metal feed pipe a meter above and connected to a tributary of a river (<http://www.telegraph.co.uk/earth/earthnews/3318094/Photographer-captures-trouts-great-escape.html>). Ideally, land-based recirculation systems would reduce the potential for escapes. However, most recirculation systems have at least partial water exchange [9] and where there is water exchange and discharge, there is a potential for escapees. These systems are still not widely used and to the authors knowledge there has not been any initiative taken to document escapees, or lack thereof, within these systems. It may, therefore, be premature to classify such systems as “escape proof.” It is unlikely that any land-based aquaculture operations could ever be 100% “escapee-proof” and, consequently, they will also need to develop anti-escapee strategies (avoiding flood plains, electric fences, grids of the appropriate mesh, catchment basins, etc.).

Moving to land-based or closed-containment operations is one approach that may help address some sustainability issues but is not without its problems. Large amounts of energy, often diesel or electric power, are required to pump and aerate water. Nutrients are

either pumped back into the water or settled somewhere and “trucked” offsite. All of these processes leave a “carbon footprint,” and only partly solve the issue of excess nutrients. IMTA, or its variation called “aquaponics,” will have to be added to closed-containment or land-based systems to treat the effluents. One “impact” may simply be traded for another. Ayer and Tyedmers [10], in their life cycle assessment of alternative aquaculture technologies, warned that one could be in a case of environmental problem shifting, not solving, where, while reducing local ecological impacts, the increase in material and energy demands may result in significant increased contributions to several environmental impacts of global concern, including global warming, nonrenewable resource depletion, and acidification.

Land-based or closed-containment operations have also been advocated as a way of controlling diseases and their transmission. However, the proponents very often equate diseases to the sole problem of sea lice, leaving the issues related to viral or bacterial pathogens unaddressed. Some concerns have been expressed that multiple species on the site might increase the risk for disease transmission. It must, however, be realized that sites in the ocean and on land will always have additional unintended species associated with the operation, ranging from microorganisms to marine mammals, depending on the situation. The question is not whether to have only one species on the site, but at what density do negative interactions occur with the unintended ones and whether there are any positive interactions associated with more diversified systems. In fact, two studies [11; Robinson, pers. comm.] have demonstrated in laboratory experiments that the blue mussel, *Mytilus edulis*, is capable of inactivating the infectious salmon anemia virus (ISAV), as well as the infectious pancreatic necrosis virus (IPNV). Mussels are, consequently, not a likely reservoir host or vector for ISAV and IPNV. Put in an IMTA perspective, this could mean that mussel rafts could be strategically placed to serve as a kind of sanitary/biosecurity cordon around salmon cages to combat certain diseases. Pang et al. [12, 13] also reported reduced total bacteria and *Vibrio* counts in a seaweed-abalone IMTA system.

In regard to parasites, two studies [14; Robinson, pers. comm.] indicate that blue mussels can consume copepodids, the planktonic and infectious stage of sea

lice, and several studies, in both Europe and New Zealand, have highlighted the fact that mussels can consume small zooplankton. Having a biofilter such as mussels at IMTA sites may decrease the frequency of exposure to pathogens and planktonic parasites. The hope is that having multiple species on a farm will result in some positive interactions between species allowing some biological control of the outbreaks of pathogens and parasites, hence reducing the number of costly chemical treatments required. If this is validated, filter feeders may have additional contributions to sustainability beyond reduction of the particle load. One of the 14 projects of the recently created Canadian Integrated Multi-Trophic Aquaculture Network (CIMTAN) is investigating the role of bivalves in potentially reducing sea lice populations. Most of the work has been conducted in the lab so far, but results are very positive and it has been demonstrated that mussels eat the larval forms. Ongoing work is developing a trap system that exploits various behaviors of sea lice to attract and filter them out of the system. Another CIMTAN project is looking into the possibility that mussels could reduce the horizontal transmission of *Loma salmonae*, responsible for microsporidial gill disease of salmon (MGDS), a serious endemic gill disorder in marine netpen reared, and wild, Chinook (and other Pacific) salmon. Trials will examine the proof of principle that blue mussels remove microsporidial spores from water and to what extent these spores retain short-term infectious potential as determined by branchial xenoma expression in test fish.

IMTA is not entering directly the debate regarding the inclusion of fishmeal and fish oil in commercial feeds (nor are land-based or closed-containment operations). IMTA could, however, provide a partial solution. Modern commercial salmon diets in Canada contain much less fishmeal (about 15–25%) and fish oil (about 15–20%) than they did less than 10 years ago (40–60%). By-products (trimmings, offal) of wild catch fisheries are now used to supply a major portion of the fishmeal ingredients. Finding replacements for marine ingredients is a priority and there are several large research projects worldwide addressing this issue. The feed company Skretting has now produced a salmon feed which includes no marine ingredients. Turning toward land plant proteins is not without its impacts. Extra farmland area (more deforestation)

would be needed, which, moreover, would need to be irrigated and fertilized on a planet already suffering from water availability problems and with fertilizer prices soaring. The price of some staple food crops used in traditional agriculture (corn, soya bean, sugar cane, etc.) would rise considerably due to announced competition for their uses, as recently seen when they were potentially sought out as energy crops for the production of first-generation biofuels [15–17]. Reallocation of acreage for subsidized potential biofuel crops such as corn, sugar cane, oil palm, canola, switch grass, etc., has already had significant ecological and societal costs due to its impacts on ecosystem health, biodiversity, and food security [18–21]. Partial substitution with organisms already living in water and not needing extra fertilization in an IMTA setting, such as seaweeds, could, in fact, be a very interesting option, fitting well within the sustainability and management concept of IMTA, and representing a logical loop for companies developing an IMTA and diversification strategy. If cultivated in the water column in IMTA systems, there would, moreover, be no issue of raking natural beds of seaweeds attached to the bottom of the ocean (destruction of seafloor and impact on ecosystem functions such as nursery ground for animals).

Some environmental nongovernmental organizations arguing for fishmeal/fish oil replacement have also voiced concerns that, after all, marine fish should eat marine ingredients . . . obviously, one cannot have it both ways! There is also the paradoxical situation of farmed freshwater fish, which are being grown less and less on humans and animal wastes and naturally occurring algal blooms, but more and more on already competing staple foods such as corn and soy: they have lost their off-flavored or muddy taste to become tasteless or “unfishy”! So, what does one want to receive in one’s kitchen? A flavor-neutral, versatile product easily adapted with numerous sauces, while one is lamenting that farmed salmon or bass are not what wild salmon or bass used to be? Quite an irony, even more so when people learn that these herbivorous whitefish are more and more being fed pellets containing fishmeal and fish oil because they grow faster! What is really important is a balanced diet using balanced sourcing of raw material.

Some will argue that “fish require nutrients, not ingredients.” At the same time, there is also the

well-known saying “You are what you eat,” and in this case, people have to realize and accept that humans are mostly corn, soya, and fishmeal, if they look at what the four mammals (cow, pig, sheep, and goat) and four poultry (chicken, turkey, duck, and goose) that they have selected as their meat choices are eating. Historically, most of the reduction fishery (small fish such as anchovies, herring, sardines, and menhaden) went into the production of pet feeds and farm animal feeds. Subsequently, this fishery supplied a significant part of the marine ingredients for fish feeds. The landing of the reduction fishery has been fairly stable (fluctuating between 15 and 30 million metric tons since the 1970s) and, in the absence of aquaculture, the fishery would likely return to supplying pet and farm animal feeds, and a current resurgence of interest directly by humans. This is not to justify relaxed vigilance for finding replacements for marine ingredients in fish feed, but simply to suggest that an absence of fish farming will not stop the use of this resource. How can one get out of this vicious circle? Cultivating several organisms, at different trophic levels, in proximity so that the food and wastes are utilized efficiently more than once through a cascade of recapturing and remetabolizing is one approach: that is IMTA. The other is to consider that if terrestrial food production systems are close to their limits, one does not have other options but to turn again to the sea, this time not for fish but to have seaweeds and invertebrates entering one’s food habits, either directly or delivered through feed given to intermediates to what reach one’s plate. The discrepancy between the marine and agricultural production systems has to be reduced: presently, especially in the western world, humans feed approximately two steps higher in the marine food web than in the agricultural food web.

People should continue to eat seafood (fish but also invertebrates and seaweeds), not according to seafood pocket guides which simplistically paint species with one stroke of green (best choice), orange (good alternative), or red (avoid), but according to the fishing and aquaculture practices used to grow, harvest, and process them: an admittedly more complex mosaic, but also much more realistic and attractive to look at than a traffic light!

Interestingly, what is referred to as the fifth tasting sense by Japanese (after sweet, sour, salty, and bitter)

and called umami (= savoriness or good flavor) comes from seaweeds. The product responsible for umami was first identified in 1908 by Professor Kikunae Ikeda, of the Tokyo Imperial University, searching for the chemical reason of the strong flavor in seaweed broth (mostly of the kelp *Saccharina japonica*, formerly *Laminaria japonica*). It is due to the detection in our mouth of the carboxylate anion of the amino acid called glutamic acid and its salts, glutamates, in particular monosodium glutamate (MSG). Inosine monophosphate (IMP) and guanosine monophosphate (GMP), degradation products of the energy-storing molecule adenosine triphosphate (ATP) greatly enhance the perceived intensity of umami. This explains, chemistry displacing romantics, why a dead tuna (once full of energy) served with seaweeds is such a savory delicacy, the very essence of the success of the sushi bar fad gaining the western world.

We have never pretended that IMTA is the solution, the silver bullet, to and for everything. It is now up to us to develop the better aquaculture practices of tomorrow. IMTA is based on several common sense principles:

- The solution to nutrification is not dilution, but extraction and conversion through diversification.
- This is, in fact, a rewording of the first law of thermodynamics “Rien ne se perd, rien ne se crée, tout se transforme” (“Nothing is lost, nothing is created, everything is transformed”) as summarized by Antoine Laurent de Lavoisier, the well-known French chemist and physicist (but also tax collector, which explains his premature death at age 51 in 1794 under the Terror period of the French Revolution).
- What is waste for some is gold for others.
- Do not put all your salmon eggs in the same basket.

A lot of common sense, but, unfortunately, common sense is not that common! IMTA is one of the promising options, but, certainly, it needs to be tailored to the location in the world where it is implemented. It should also be developed in association with other practices. Like for energy, not one solution will satisfy all the needs and it is a variety of solutions that will help one secure one’s seafood procurement in a responsible manner. The solutions will be at the interfaces of these techniques and will be interdisciplinary. They will



embrace both scientific and technological advancements and traditional knowledge. IMTA is exactly at this interface, modernizing traditional practices: combining ecosystem complementary crops, bay management area, and fallowing are nothing new, but revisited and updated, based on what humans have learned from past experience (which includes a lot of mistakes over the centuries, but not assimilated by the characteristically short-term memory of humans!).

### **Recognizing and Valuing the Biomitigative Services Rendered by the Extractive Components of IMTA: Should a System of Nutrient and Carbon Trading Credits Be Developed?**

A few economic analyses have indicated that the outlook for increased profitability through IMTA is promising [22, 23]. However, these analyses were based solely on the commercial values from the sale of biomass – being of fish, shellfish, or seaweeds – and used conservative price estimates for the cocultivated organisms based on known applications. One aspect not factored into these analyses is the fact that the extractive component of an IMTA system not only produces a valuable multipurpose biomass, but also simultaneously renders waste reduction services to society. It is particularly important to recognize that once nutrients have entered coastal ecosystems, there are not many removal options available: the use of extractive species is one of the few realistic and cost-effective options. The economic values of the environmental/societal services of extractive species should, therefore, be recognized and accounted for in the evaluation of the true value of the IMTA components. Further development of economic models is needed to help shed light on the economic (society) and commercial (industry) attractiveness of IMTA.

Ecosystem services have been ignored until recently [24]. To improve the sustainability of anthropogenic nutrient loading practices such as aquaculture, incentives such as Nutrient Trading Credits (NTC) should be established as a means to promote nutrient load reduction or nutrient recovery. During the last few years, there has been much talk and excitement about carbon credits. However, within coastal settings, the concerns have largely been with nitrogen, due to the fact that its typical role as the limiting nutrient is not any longer the

case in some regions. Potential effects of carbon loading in the marine environment should also be considered: localized benthic anoxia and, consequently, hydrogen sulfide release may occur when solid waste deposition rate exceeds aerobic decomposition rate. Ocean acidification due to increased dissolved CO<sub>2</sub> levels has also prompted serious new concerns [25] and a Carbon Trading Credit (CTC) system should also be contemplated. With an appropriate composition of cocultured species, IMTA has the potential to reduce the amounts of dissolved (inorganic) and solid (organic) forms of nitrogen, carbon, phosphorus (more an issue in freshwater environments), etc., making extractive aquaculture a good candidate for a NTC and CTC, or other suitable approaches, to deal with the pressing issues of coastal nutrient loading.

Currently, there are few countries with laws or regulations that require aquaculture operations to responsibly internalize their environmental costs, such as nutrient discharges. There are some precedents, such as where land-based trout farmers in Denmark are allowed to increase their feed quota with documented evidence of reduced effluent discharge [26], but such incentives are not widely spread. In most jurisdictions, adjacent ecosystems are left to accommodate the nutrient load, and performance-based standards are used to determine if farms have exceeded their assimilative capacity.

The implementation of regulations resulting in internalization of environment costs by fish farms, without a direct economic compensatory response such as the Danish feed quota increase, could result in a significant reduction in profitability. In land-based systems, it is relatively easy to quantify nutrient load and concentration via comparison between farm inflows and outflows, thereby creating a benchmark for “economic compensation.” Such values are practically impossible to empirically measure in an open-water system, “leaky” by definition, and, consequently, so is the practical implementation of such incentives. However, Troell et al. [27] and Chopin et al. [28] demonstrated that by integrating the seaweed, *Gracilaria*, in the dual role of nutrient scrubber and commercial crop (for agar production), with salmon farms in Chile, the environmental costs of waste discharges would be significantly reduced and profitability significantly increased.

Interestingly, the removal of nitrogen could be much more lucrative, by approximately a factor of 100, than that of carbon. The cost of removing nitrogen is not clearly defined, but there are several interesting studies that may help define a range of possible prices for economic evaluation of the NTC concept. The cost of removing 1 kg of nitrogen varies between US\$3 and US\$38 at sewage treatment facilities, depending on the technology used and the labor costs in different countries [28]. The municipality of Lysekil, in Sweden, is paying approximately US\$10/kg removed by the filter-feeding mussel, *Mytilus edulis*, to the farm Nordic Shell Produktion AB [29, 30]. Ferreira et al. [31, 32], with the development of the Farm Aquaculture Resource Management (FARM) model, determined a net value of €18–26 billion/year of nutrient eutrophication reduction services provided by shellfish aquaculture in the coastal waters of the European Union. Gren et al. [33] calculated that the cleaning costs of nutrients by mussel farming can be considerably lower than other abatement measures and estimated that mussel farming should be credited between €0.1 and €1.1 billion/year in the Baltic Sea.

Using this information, and without presuming what the final design of IMTA sites will be in the future, preliminary calculations for the relatively small-scale

IMTA project on the East coast of Canada indicate that the annual harvesting of kelps (Fig. 3) would equate to the removal of 35.75 t of nitrogen from the ecosystem, representing an NTC of between US\$357,504 and US\$1,072, 512. The same could be applied to another key nutrient, phosphorus. With an annual removal of 4.09 t and a value of US\$4/kg removed [28], this would represent another contribution to the NTC of US\$16,343, a much smaller amount but it could also be an important way of extracting phosphorus, at a time when some are predicting it to be the next element human society will be short of (in its natural or mined forms).

Carbon Trading Credits (CTC) could also be calculated. There may be some arguments about what is meant by trapping and sequestering carbon. Some may argue that it should be reserved to long/geological term storage (sink) and not to transient storage [34]. This is, in fact, a question of how long one allows the recycling clock to run. There is no permanent storage of carbon; it happened that a particular fossil biofuel, petroleum, has been sequestered over geological time to suddenly be reused at an accelerated rate over the last few centuries. But the first law of thermodynamics, as enunciated by Antoine Laurent de Lavoisier more than two centuries ago, still applies: “Rien ne se perd, rien ne



**Aquaculture, Integrated Multi-trophic (IMTA). Figure 3**

Harvesting of the kelp, *Saccharina latissima*, at an Integrated Multi-Trophic Aquaculture (IMTA) site in the Bay of Fundy, New Brunswick, Canada. Kelps remove dissolved nutrients from the ecosystem while providing commercial products

se crée, tout se transforme,” i.e., “Nothing is lost, nothing is created, everything is transformed.” If even temporary removal of carbon from the ocean by biomass harvesting until further transformation (and rerelease of carbon) can be credited for potentially increasing seawater pH and absorbing CO<sub>2</sub> from the atmosphere and/or the cultivated animals, then CTC should be calculated. Marine vegetation is getting more and more recognition as a sink for anthropogenic carbon emissions (the so-called blue carbon [35]). Marine primary producers contribute at least 50% of the world’s carbon fixation and may account for as much as 71% of all carbon storage in oceanic sediments. Then, micro-algae, macro-algae, and marine plants, such as mangroves and seagrasses, have a role to play in CO<sub>2</sub> sequestration and removal, and carbon storage [36]. Marine photosynthesis accounts for 50% of the total primary productivity of the planet (54–59 PgC/year from a total of 111–117 PgC/year [37]). Of this, marine macrophytes (seaweeds and seagrasses) account for approximately 1 PgC/year concentrated in coastal regions where they can play a significant role in the sequestration of anthropogenic carbon emissions and the global carbon cycle. Brown marine macro-algae (such as *Macrocystis*, *Saccharina*, *Laminaria*, *Ecklonia*, *Sargassum*, *Ascophyllum*, and *Fucus*), red algae (such as *Porphyra*, *Palmaria*, *Eucheuma* and *Gracilaria*) and green algae (such as *Ulva*), are capable of very high rates of photosynthesis and productivity. These rates of productivity compare very favorably to those of terrestrial crops that have been recommended as possible sources of first-generation biofuels (corn, *Zea mays*) or second-generation biofuels (switch grass, *Panicum virgatus*; E-grass, *Miscanthus giganteus*) and position marine macro-algae very well for being part of the third-generation biofuels [36].

Coming back to the IMTA project on the East coast of Canada, using a value for carbon removal of around US\$30/t [34], the annual harvesting of kelps would represent an annual removal of 306.43 t and a CTC of US\$9,193: a larger amount of carbon, but for a much smaller value of trading credits, underlining the difficulty in removing dissolved nutrients from aquatic systems and the acute issue of their presence in coastal systems. Similar calculations could be applied to the organic extractive component of IMTA. In the case of

shellfish, accumulation of nitrogen, phosphorus, and carbon should be considered both in meat and shells, which are especially rich in calcium carbonates.

At a much larger scale, the occurrence of large and recurrent “green tides” should also be brought into focus. Large proliferations of opportunistic green algae, especially of the genus *Ulva*, in response to large anthropogenic nutrient loading, have been in the news over the last few years in places around the world such as Northern Brittany in France, the southern regions of the UK, and Venice in Italy. The green tide in Qingdao, China, just before the sailing competitions of the 2008 Olympic Games, got a lot of attention (Fig. 4). The following question needs to be asked: Are these green tides a negative media photo opportunity, or are they reminders of the significant role seaweeds play in coastal processes and the services they render? Within 3 weeks, 1 million tons of *Ulva prolifera* were removed from the vicinity of Qingdao to allow the sailors and windsurfers to compete (but it is estimated that approximately 2 million tons of *U. prolifera* sank to the bottom of the Bay; another environmental problem shifting, but not a solution). The harvesting of 1 million tons equated to between 3,000 and 5,000 t of nitrogen removal for a NTC value between US\$30 and US\$150 millions! Additional NTC of US\$1.6 million for the removal of 400 t of phosphorus, and CTC of US\$900,000 for the removal of 30,000 t of carbon, should also be factored in.

A smaller green tide occurred in 2007. Large ones were also reported in 2009 and 2010 but they stayed offshore in the Yellow Sea [38, 39]. Out of sight should, however, not mean out of mind. If urgent measures are not taken, this will be a recurrent event for years to come. Is there a solution? Green tides are not the cause, but the unintentional consequence of coastal eutrophication. With the presence of sufficient nutrients and solar energy, these opportunistic species, with a well-adapted anatomy, morphology, and physiology, will proliferate. Obviously, it would be beneficial to reduce nutrient loading at the source, but this may not be possible in the present context of economic development along China’s coastal zone. The problem is that *U. prolifera* is presently an unwanted and uncontrolled growing nuisance species of limited commercial value. To control its proliferation, the solution may be to create a competition for nutrients by intentionally



**Aquaculture, Integrated Multi-trophic (IMTA). Figure 4**

A green tide of *Ulva prolifera* in Qingdao, China, just before the 2008 Olympic Games, triggered a massive cleanup

cultivating algal species, which not only carry on the biomitigation, but also have a commercial value, where *U. prolifera* starts to enter the coastal environment (discharges from juvenile river crab land-based aquaculture ponds along Jiangsu province, south of Shandong province where Qingdao is located). This time, the IMTA concept has to be interpreted as an integrated land pond/coastal aquaculture system in a supra Integrated Coastal Zone Management (ICZM) effort, beyond provincial borders, to address issues at the Yellow Sea scale. It is understood that this “out of the box” approach to ICZM will, initially, raise eyebrows as the idea of growing more seaweeds (but of commercial value) to contain the proliferation of other seaweeds, presently considered nuisances, is not the most intuitive approach for a lot of people or decision makers! The question is simple: what are the best nutrient scrubbers once nutrients are in a dissolved state and have reached coastal waters? The answer is seaweeds, but can people, preferably, grow the ones they have applications for?

At the present time, there seems to be a stage of recognition, awareness, and communication of the concepts of ecosystem services and biomitigative services rendered by extractive aquaculture (the differences between the two not always being clearly

identified and explained in some publications). Next will come the time to transform the concepts into biomitigative solutions and then their inclusion in regulatory and management frameworks. Establishing and implementing a structure for the payment schemes (credits or incentives) of these services will be a delicate matter. Will it be one agency, but with funds coming from where? Should it be a regional, national, or international agency(ies), trading at which scale(s)? Will an extractive aquaculture operation in existence for many years receive credits, or will only the new ones? Would a fed aquaculture operation also practicing extractive aquaculture be eligible for credits, or will it be the case for the extractive only aquaculture operations? What about the situation in which people run both types of farms. Moreover, due to complex hydrographic and current patterns, it is obvious that extractive species at a site are not limited to absorbing/sequestering the nutrients generated exclusively at that site. Consequently, is it possible to establish a clear spatial nutrient removal budget which would be associated with the corresponding credits/incentives? Will the sequestration have to be “permanent,” or will a temporary removal/storage be acceptable and more realistic? A lot of regulatory details will have to be worked out before this complex scheme becomes a reality.



### What Will It Take to Increase the Acceptance and Adoption of IMTA as a Responsible Aquaculture Practice of the Future?

Presently, the most advanced IMTA systems in open marine waters and land-based operations have three components (fish, suspension feeders or grazers such as shellfish, and seaweeds, in cages, rafts, or floating lines), but they are admittedly simplified systems [40]. More advanced systems will have several other components (e.g., crustaceans in mid-water reefs; deposit feeders such as sea cucumbers, sea urchins and polychaetes in bottom cages or suspended trays; and bottom-dwelling fish in bottom cages) to perform either different or similar functions, but for various size ranges of particles, or selected for their presence at different times of the year (e.g., different species of seaweeds). The most advanced IMTA systems, near or at commercial scale, can be found in Canada, Chile, South Africa, Israel, and China [41, 42]. Ongoing research projects related to the development of IMTA are taking place in the UK (mostly Scotland), Ireland, Spain, Portugal, France, Turkey, Norway, Japan, Korea, Thailand, the USA, and Mexico. It will also be interesting to observe how new seaweed cultivation initiatives in different parts of the world for biofuel production could be an additional driver to adopt IMTA practices.

For IMTA to develop to a commercial scale, appropriate regulatory and policy frameworks need to be put in place. Present aquaculture regulations and policies are often inherited from previous fishery frameworks and reasoning, which have shown their limitations. It is, therefore possible that some of the existing regulations and policies could impose unintentional constraints on the future growth of IMTA. To develop the aquaculture of tomorrow, current governance structures pertaining to aquaculture need to be revisited and reviewed with the aim of identifying changes in the regulatory/policy environment that are needed to facilitate the operation of IMTA farms. Adaptive regulations need to be developed by regulators with flexible and innovative minds, who are not afraid to put in place mechanisms that allow the testing of innovative practices at the R&D level, and, if deemed promising, mechanisms that will take these practices all the way to C (commercialization). As the IMTA concept continues to evolve, it is important that all sectors of the

industry are aware of the implications of the changes involved, so that they can adapt in a timely and organized manner.

To move research from the “pilot” scale to the “scale up” stage, some current regulations and policies may need to be changed or they will be seen as impediments by industrial partners who will see no incentive in developing IMTA. For example, an earlier version of the Canadian Shellfish Sanitation Program (CSSP) prevented the development of IMTA because of a clause that specified that shellfish could not be grown closer than 125 m of finfish netpens. This paragraph was clearly not written with IMTA in mind, but it seriously impinged its development. After 4 years (2004–2008), it was amended so that IMTA practices could develop to commercial scale legally, based on recent, reliable, and relevant data and information provided by three government departments and the IMTA project on the east coast of Canada. While 4 years may seem long, it is a relatively short delay considering that regulations and legislations require thorough review with due governmental process involving several federal and provincial departments. This suggests that new aquaculture practices should be accompanied by timely regulatory review to avoid market delays for new products. As governments move to revise current regulatory regimes, it will be necessary to press the importance of accommodating and indeed encouraging new sustainable solutions such as IMTA. IMTA also requires approaching aquaculture development and management with a holistic approach and not one species, or group of species, at a time. It is known that this approach has led to many failures in the management of the fisheries; vigilance is required so that the same flaw is not repeated in the management of aquaculture.

Most current aquaculture business models do not consider or recognize the economic value of the biomitigative services provided by biofilters, as there is often no cost associated with aquaculture discharges/effluents in land-based or open-water systems. In order to ensure further development of IMTA systems worldwide, from the experimental concept to the full commercial scale, defining and implementing the appropriate regulatory and policy frameworks, and financial incentive tools such as NTC and CTC, may therefore be required to clearly recognize the benefits of the extractive components of IMTA systems. Better



estimates of the overall costs and benefits to nature and society of aquaculture waste and its mitigation would create powerful financial and regulatory incentives to governments and the industry to jointly invest in the IMTA approach, as the economic demonstration of its validity would be even more obvious. Moreover, by implementing better management practices, the aquaculture industry should increase its societal acceptability, a variable to which it is very difficult to give a monetary value, but an imperative condition for the development of its full potential. Reducing environmental and economic risk in the long term should also make financing easier to obtain from banking institutions [43].

The determination to develop IMTA systems will, however, only come about if there are some visionary changes in political, social, and economic reasoning. This will be accomplished by seeking sustainability, long-term profitability, and responsible management of coastal waters. There is still a large amount of education required to bring society into the mindset of incorporating IMTA into their suite of social values. Some of the attitudinal surveys conducted in Canada [23, 44] and the USA [45] indicate that the general public is in favor of practices based on the “recycling concept.” Consumers’ perceptions and attitudes may also have to change. Why is recycling and the concept of “what is waste for some is gold for others” well accepted in agricultural practices, but is not yet acquired when transposed to aquaculture practices? Will consumers come to accept eating products cultured in the marine environment in the same way they accept eating products from recycling and organic agricultural practices, for which they are willing to pay a higher price for the perceived higher quality or ethical premiums? After all, regulations require mushrooms to be specifically grown on farmyard manure and animal excrements to receive organic certification (European Community Regulations No 2008R0889 – Article 6). Will a greater appreciation of the sustainable ecological value of the IMTA concept, a willingness to support it tangibly with shopping money, and an increased pressure on elected representatives emerge? This will be the ultimate test. The degree to which researchers and extension people become creatively involved with this educational component will be vital to the success of IMTA practices. The differentiation of IMTA products through

traceability and eco-labeling will also be key for their recognition and command of premium market prices.

Some have argued that the adoption of IMTA in the western world is slow. For example, on the east coast of Canada, there were obviously no IMTA sites in the Bay of Fundy in 2001 when IMTA research started. Nine years later, 8 of the 96 finfish sites in southwest New Brunswick have the combination salmon (or cod)/mussels/kelps and 8 other sites have been amended to develop IMTA. This is a respectable conversion of almost 16% in 9 years. Moreover, it would not be reasonable to anticipate an instant conversion, as the industry needs to develop markets to absorb the cocultured biomass: this also takes time and can only be progressive.

### Future Directions: The Path Forward

Several IMTA projects, worldwide, have now accumulated enough data to support the proof of concept at the biological level. The next step is the scaling up of more experimental systems to commercial scale to further document the economic and social advantages of the concept, which will be key to offering IMTA to practitioners of monospecific aquaculture as a viable option to their current practices. Emerging responsible aquaculture approaches must generate net economic benefits for society if they are to be advocated. Working on appropriate food safety regulatory and policy frameworks in the respective countries will be essential for enabling the development of commercial scale IMTA operations in a more universal fashion.

It has taken decades to reach current finfish aquaculture production levels and learn new species husbandry. A major rethinking is, however, needed regarding the definition of an “aquaculture farm” by reinterpreting the notion of site-lease areas and regarding how it works within an ecosystem, in the context of a broader framework. Within Integrated Coastal Zone Management (ICZM), integration can range from the small scale (a leased site with its spatial limits) to a Bay Management Area (BMA) and to the larger scale of a region connected by the functionalities of the ecosystem. Amending regulations to allow a new type of aquaculture systems will not occur overnight. This should, however, not discourage the finfish aquaculture industry from practicing IMTA, as even small amounts

of cocultured species production are useful at the initial stage of development.

Selecting the right combination of species will be critical. They will have to be appropriate for the habitat, the available culture technologies and labor forces, and the environmental, climatic, and oceanographic conditions. They will have to be complementary in their ecosystem functions, growing to a significant biomass for efficient biomitigation, commanding an interesting price as raw material or presenting an interesting added value for their derived products. Their ecological interactions and synergies within an IMTA system will have to be identified and understood to take full advantage of them. Their commercialization should not generate insurmountable regulatory hurdles.

Optimal design will not only facilitate nutrient recovery, but should also promote augmented growth beyond what would be expected were these species cultured in isolation. In addition to the obvious economic return from increased growth rates from additional species, some less tangible benefits should also be factored in, such as the biomitigative services rendered by the extractive species. Economic analyses will have to recognize and account for the values of the environmental/societal services of extractive crops to estimate the true value of these IMTA components. Economic analyses will need to be part of the overall modelling of IMTA systems, as they get closer to commercial scale and their economic benefits and costs, as well as their impacts on coastal communities, are better understood. It will then be possible to add profitability, resilience, social/economic desirability, and economic impacts to the comparison between IMTA and monoculture settings. They will have to include the pricing and marketing potential and impact of organic and other eco-labellings, the value of biomitigative services for enhanced ecosystem resilience, the savings due to multi-trophic conversion of feed and energy which would otherwise be lost, and the reduction of risks through crop diversification and increased societal acceptability of aquaculture (including food safety, food security, and consumer attitudes toward buying sustainable seafood products). This would create economic incentives to encourage aquaculturists to further develop and implement sustainable marine agronomy practices such as IMTA, and would increase the societal acceptability of aquaculture by the general

public. Seaweeds and invertebrates produced in IMTA systems should be considered as candidates for a variety of regulatory measures that internalize these benefits. For example, nutrient and carbon trading credits (NTC and CTC) could be used to promote nutrient removal, CO<sub>2</sub> sequestration, oxygen provision, and coastal eutrophication reduction within the broader context of ecosystem goods and services. Long-term planning/zoning promoting biomitigative solutions, such as IMTA, should become an integral part of coastal regulatory and management frameworks.

Nutrient extractive aquaculture appears to be a viable ecological engineering option for managing/internalizing some of the externalities generated by aquaculture operations. Effective government legislation/regulations and incentives to facilitate the development of IMTA practices and the commercialization of IMTA products will be necessary. The development and adoption of technology often depends in part on the level of legislative pressure from a nation's government, itself reacting to pressures from consumers, environmental nongovernmental organizations, and the public at large. If environmental legislation remains a low priority with government, then little progress toward the use of biofilters (as a means of effluent mitigation) will occur. The only motivator will be profits obtained from additional product growth and regulatory incentives. Therefore, if governments put legislative pressure on the proper management of wastewater effluent, openly support the use of biomitigation for effluent management, and put in place the appropriate corresponding financial tools (funding for IMTA Research & Development, outreach and technology transfer, and NTC and CTC incentives), then the development of IMTA will be encouraged.

### **Caution: Let's Not Promise the Moon and Let's Be Conscious of Societal Constraints, Particularly in the Western World**

During the last few years, there has been a renewed interest in the mariculture of seaweeds and their uses, something that should make phycologists and ecologists rejoice, as this group of organisms, never clearly systematically circumscribed, has been misunderstood, unappreciated and under/misused over the centuries.

There is now an opportunity to explain what seaweeds are, and the many applications, benefits, and services they can provide. However, how can people do that appropriately and responsibly, without “promising the moon” that they will not necessarily attain, and risking another “purgatory period” in between each energy crisis?! Seaweeds (and algae in general) made the news in the 1970s–1980s; they are back in the news now (2000s–2010s). If people are not careful to distance themselves from charlatanistic claims, which abound in the media and even in certain scientific circles, they could be in a situation of not developing a sustained public interest and use of these organisms, but be in another phase of denial until the next fad cycle (2030s–2040s?), which is not productive for the acquisition of still much needed scientific knowledge, nor the teaching of our discipline or the placement of our in-between fashion students. While everyone wants the seaweed sector to develop, some biotechnological issues and societal constraints, particularly in the Western World, should be recognized and a responsible and gradual implementation strategy for the long term should be adopted.

The western marine biology community has been dominated by people who have received a mostly zoological training from kindergarten to high school, very often reinforced by a monospecific (or monogrouping) specialization at university, instead of receiving and developing an ecosystem approach to knowledge and issue solving, which are then sadly missing when concepts of ecosystem-based management, species cocultivation, and interdisciplinarity are mentioned. Not surprisingly, the knowledge of seaweeds and their functions and services in/to the ecosystem is reduced and remains at universities and research institutions that have been wise in keeping their diverse expertise, instead of succumbing to fad cycles, which, then, force them to periodically reinvent the wheel. The consequence is that every time one wants to raise the possibility of using seaweeds in research and development and commercialization (R&D&C), one has to go through a lonely period of “preaching in the desert” before facts and common sense start to prevail.

One key, common, and deeply rooted misunderstanding to shake from the minds of people is that there is more than fish in the ocean! Over the centuries humans have been quite minimalist in their meat

choices: four mammals (cow, pig, sheep, and goat) and four poultry (chicken, turkey, duck, and goose), hence, Paul Greenberg’s idea of four fish (salmon, sea bass, cod, and tuna) for the title of his book [8]. However, the ocean cannot function with only fish, and the seafood solutions cannot come from within only this group of organisms. Maybe the problem resides deeply among the English-speaking people with this overuse of “fish”: fish is a noun, which can even encompass shellfish and seaweeds in its general use, and fish is a verb. . . if you go harvest seaweeds along the shore you could be paradoxically fishing seaweeds, which for a Cartesian French-speaking person does not make much sense! In French, there is “poisson” as a noun and “pêcher” as a verb, even if both come from “piscis” in Latin. So, when a French person “va à la pêche,” it is not necessarily to get a fish, but also to go “à la pêche aux moules” (mussels), “aux oursins” (sea urchins), or “aux algues marines” (seaweeds, for which many languages also have a higher opinion, as marine algae, instead of weeds of the sea!). To function, IMTA requires, in fact, not four components but five: the fed organisms (e.g., fish or shrimps), the extractive inorganic component (e.g., seaweeds or other aquatic vegetation), the extractive small organic component (e.g., suspension feeders such as shellfish), the extractive large organic component (e.g., deposit feeders such as sea-urchins, sea cucumbers, or sea worms), and certainly a fifth component, the microbial component, of which presently not much is known. So, if people want aquaculture to work, they have to stop being obsessed with fish aquaculture! Paradoxically, it is interesting to know that fish aquaculture, of which so much is heard, represents, in fact, only 9% of the total mariculture (aquaculture in the marine environment). Shellfish aquaculture represents 43%. Seaweed aquaculture represents even more (46%), but 99.8% of it is carried out in Asia, hence the ignorance in the western world [46, 47].

It is also important to understand that sustained successful ventures rarely happen overnight and that more than a 3 year grant is generally necessary to successfully take a concept along the R&D&C continuum. For example, the IMTA program on the east coast of Canada is starting to collect the fruits of its tireless efforts as it enters its 15th year of activities, which so far could be divided into four periods: (1) the “preaching

in the desert” period from 1995 to 2000 [48], (2) the R&D proof of concept period from 2001 to 2006, (3) the R&D&C pilot scale period from 2006 to 2012, overlapping with (4) the R&D&C industrial-scale and networking period with the establishment of CIMTAN since 2009. People, consequently, have to stay away from claims of solving hunger in the world, converting everybody into frequent direct “seaweedivores,” 100% biomitigation (which, in fact, is not necessarily the goal), renewing energy at unbelievable rates that defy the rules and equations governing photosynthesis, and all that within the next 5 years with the almighty, miraculous seaweeds and micro-algae!

If there is no shortage of interesting ideas that work at the small demonstration scale, the problems generally appear when scaling up is contemplated and people start to realize what the consequences will be and, especially, the realistic, or unrealistic, deployment footprints required to implement these experimental ideas to commercial-market scales, which should make sense from environmental, economic, and production perspectives and also have an acceptable societal impact.

People should also stay away from the cliché that around 71% of this planet is covered by oceans and that, consequently, there is a lot of space for aquaculture development. If aquaculture will most probably expand into more exposed and open ocean locations in the future, due to the reduced availability of new and appropriate sheltered nearshore sites, it is doubtful that one will see farms in the middle of the Atlantic, Pacific, and Indian Oceans, due to simple logistics and weather issues. Moreover, the present international law of the sea is not that comforting for privately owned equipment (farms in this case) found at sea. The vagueness of territorial jurisdictional competence in the Exclusive Economic Zone (EEZ) in different countries, and certainly in international waters, has been a major impediment to progress of the so-called offshore aquaculture. Moving to the open ocean has been considered a means for moving away from environmental and public perception issues in the coastal zone. However, this move should not encourage an “out of sight, out of mind” attitude, as open ocean development will also come under scrutiny by a more and more educated public. Even if greater residual currents, deeper waters, and lower nutrient baselines are expected to reduce impacts from open ocean operations through wider

dispersion plumes of nutrients, as compared to similarly sized nearshore operations, there will be a point when reasonably accessible and manageable open ocean ecosystems will eventually reach their assimilative carrying capacities. Why should one think that open ocean aquaculture, the “last frontier,” will be without its own border/limits? Despite the sea being so immense, one is now learning the hard way the concept of overfishing . . . Instead of taking the position that in open ocean environments the hydrodynamic conditions will be appropriate for dispersion (a way of exporting problems, not solving them) and reduced environmental impacts (but at a significant cost in lost food), the open ocean aquaculture sector will also have to capitalize on recapturing the by-products of fed aquaculture and, hopefully, engineer, right from the beginning, efficient open ocean IMTA systems with their built-in biomitigative functions. The solution to nutrification in the open ocean environment, like in the nearshore environment, should not be dilution, but extraction and conversion through diversification. Why repeat what was done with the development of nearshore aquaculture (fish aquaculture development in the 1970s and IMTA development in the 2000s) with open ocean aquaculture (moving the fish to the open ocean in the 2010s . . . oh, the extractive species should have also been moved in the 2050s!)? These open ocean systems will also require trophic diversification from an environmental and economic perspective, with “service species” from lower trophic levels (mainly seaweeds and invertebrates) performing ecosystem balancing functions while representing value-added crops [49, 50]. Open ocean IMTA should not be an afterthought for 2050.

For some, the ecological, engineering, economic, and social challenges remaining to be solved may be daunting. However, our goal is to develop modern IMTA systems, which are bound to play a major role worldwide in sustainable expansions of the aquaculture operations of tomorrow, within their balanced ecosystem, to respond to a worldwide increasing seafood demand with a new paradigm in the design of the most efficient food production systems. There are no simple solutions, but one thing is certain – the human population is increasing on this planet and as people get richer, and their standards of living increase, they want more meat and dairy products in their diet, the

temptation of the “western diet,” while, ironically, Westerners aspire to change their diets! Will terrestrial agriculture be able to continue to supply most of this food? A balanced and responsible diet is required, and some of this food will have to come, increasingly, from aquatic food production systems, be them in seawater, brackish water, or freshwater. As was the case on land, where the acquisition of food by hunter/gatherer societies had to evolve toward agricultural practices, humans will have to accept an evolution in seafood procurement. It has to be understood, particularly in the western world, that “the modern global supermarket has a basic internal ecology” [8]. The average consumer is not a “foodie” and is not that interested in or cannot afford local, seasonal, less-than-100-miles food if not rich enough or not living within a region graced by a clement climate year long. The modern supermarket wants guaranteed supply on a 12 month basis, with limited variability in seasonality and quality. Most of the time, agricultural products can provide that comfort, barring the occurrence of an unexpected disease, contamination, drought, flood, economic protectionism, or political barrier. The seafood counter is a much more variable department to manage, at the present time, with a convoluted succession of many intermediates before seafood arrives on ice at a supermarket. It is interesting to note that the aquaculture industry’s ability to provide 12 month availability of its products, moreover of consistent quality, is improving.

People are presently witnessing the emergence of a plethora of organizations developing their own standards and eco-label/certification schemes as they jockey for position in the global marketplace. The problem is that there are presently too many possible horses to ride and nobody really knows which one(s) will cross the finish line and, consequently, which one(s) to bet on as worth being associated with. One can only wonder what will happen when so many fisheries and aquaculture operations will be eco-certified. If everything is certified, nothing will be certified ... and certification will lose its aura the same way some argue organic labelling is losing its significance, after having been used and overused. All that, of course, to the great confusion of the consumers, who cannot follow this contradictory debate/competition among standard setters, and may decide to simply stay away from seafood all together when, in fact, seafood

products are healthy [51]. One of the problems is that some of these standards are passing or failing grades, with no incentives for continuous improvement from a minimal baseline to be decided, followed by a tiered approach. Some argue that it would give accreditation to companies at a very low level. However, putting the bar so high is not a recipe for gradual improvement of everybody involved, to progress and gradually reach the ultimate goal, although admittedly not overnight. If 20% of the global farmed seafood producers are certified at the highest threshold, what happens to the remaining 80% and the chance of incentivizing them to improve their practices? What happens when, in a bay management area, several aquaculture companies have taken the appropriate measures to be certified, but a “black sheep” (should it be a black cod?!) makes the whole certification scheme crumble once the hydrodynamics of the bay are considered? By analogy, in which the vector this time is not water but wind, one sees the same dilemma in parallel agriculture situations where conventional and organic agriculture practices are separated by illusionary buffer zones. On one hand, one can understand the desire by suppliers and retailers to see a hard to meet certification scheme so as to differentiate themselves from the others (most probably amounting to the privilege of displaying a sticker or logo on the packaging); on the other hand, too high a certification carrot, or moving goalposts, may not be the best strategy if progress toward overall better and more responsible aquaculture practices is the goal. The market will ultimately decide who remains in the competition and which logo(s) will be trusted by the general public, but there still are several years of confusion ahead.

Agricultural development has been associated with significant changes in landscape and land use; one can expect that the evolution of sourcing one’s seafood more and more through aquaculture will also trigger significant “seascape” and “sea use” modifications, all the way to one’s deepest human social structures and governance. The transformation from hunters/gatherers to farmers happened many centuries ago on land. Humans are in the middle of this transformation at sea and that is maybe why they are so uncomfortable with this evolution they are part of, and not able to sit back and analyze without being emotional. It is up to them to be a link in the chain, which will hopefully lead



to fishing and aquaculture practices done right, enabling them to become herders and farmers of the sea. It should not be forgotten that they are still in the infancy of modern, intensive aquaculture and that some agricultural practices have taken centuries to develop into better, not necessarily yet best, management practices.

Beyond the market and marketing issues and the biological, environmental, economic, technological, engineering, and regulatory issues of aquaculture developments, the basic question will be that of societal acceptance. Are humans ready to evolve in their use of the “last frontier” of this planet and consider not only the challenges of the physical forces at sea (wave exposure, winds, currents, depth, etc.) but also those of shipping routes, fishing zones, offshore gas and mineral extraction areas, migration routes for marine mammals and birds, recreational uses, and then finally deal with the concept of zoning some portions of the oceans for large aquaculture parks, as sustainable food production systems for an ever-seafood-hungry human population? Despite all the campaigns, boycotts, documentaries, books, seafood pocket guides, scare tactics, sustainable/local/seasonal movements among affluent restaurant goers in weather clement regions and western world well offs, one can only admit that the global human population continues to grow and eat more seafood than ever per capita per year. So, where does that leave people? Paul Greenberg wrote that very often people consider fish as “a crop, harvested from the sea that magically grew itself back every year. A crop that never required planting” [8]. But are they investing in the principal, being in fisheries or in aquaculture, to only harvest the interests every year so as to not reduce/eat the capital for long-term sustainability? Are people ready to put some savings aside in the form of Marine Protected Areas (MPAs), not only for their natural beauty, but also for their functions in the ecosystem such as breeding grounds, nursery habitats, and food production areas? It seems that the concept of zoning the sea, or what is now called, in a softer terminology, “marine spatial planning” (MPS), is finally starting to be legislated in some countries, notably in the UK and the USA.

The same question of readiness for marine spatial planning could also be applied to emerging projects of wind farms and biofuel farms at sea. In fact, combining

IMTA open-ocean farms with wind, underwater turbine, and/or biofuel farms into large multipurpose integrated food and renewable energy parks (IFREP) could be a means for reducing their cumulative footprint, while integrating green energy with food and fuel production and processing [52]. Our business models will have to change from “one species – one process – one product” to a streamed bioeconomic chain, or web, approach among different industry sectors for the production, on one hand, of a wide range of bio-based, high-valued food and feed products/ingredients/supplements, specialty fine and bulked chemicals, agrichemicals, biostimulants, pharmaceuticals, nutraceuticals, functional foods, cosmeceuticals, botanicals, pigments and, on the other hand, lower-valued commodity energy carrying molecules/biofuels, all of them produced within reduced footprint requirements. The synergies and the services rendered by cultivating organisms of different trophic levels in an integrated manner will have to be understood and valued. The physiological, biochemical, and production performances of the different organisms will have to be improved to make the systems even more efficient, profitable, and competitive. The aquaculturists and different multi-sector end users will need to become interdisciplinary in their approach and learn to collaborate and share/integrate the biomass cultivation and processing steps (production, harvesting, pretreatment and transportation, separation and fractionation, and sequential biomass processing), while aiming at the lowest resource and energy inputs. Culture diversification into species that might otherwise be inappropriate for food markets fits well within the sustainability and management concept of IMTA. Functionalities will have to be maintained, as much as possible, along the process for optimal use/valorization of the multipurpose biomass, and not necessarily the maximization of just one end product, as some coproducts will, in fact, reveal themselves as the real drivers of the emerging integrated sequential bio-refinery (ISBR) concept [53], extended to macro-algae instead of only considering micro-algae. Market volumes/values, biomitigative services, and public acceptance will have to be included and fit into the models.

If the “Not In My BackYard” (NIMBY) and the “Build Absolutely Nothing Anywhere Near Anything” (BANANA) attitudes continue to prevail, especially in

the western world, then humans will not be able to secure their food, chemicals, and energy in an intricately interconnected ecosystem responsible manner, despite all the rhetoric heard today regarding alternative technologies and solutions (the so-called “greenwash”). Self-sufficiency of humans will not be ensured but will become dependent on other food, chemicals, and energy “masters,” who may no longer be in the Middle East but instead in the Far East (99.8% of the 15.8 million tons of cultivated seaweeds come from China, Indonesia, the Philippines, Korea, and Japan [46, 47]). It is time to walk the talk and recognize the implications – notably regarding marine spatial planning and our societal production and food habits – of the policies elaborated for the future.

The 1960s were the time of the “Green Revolution” on land, but some would question if it was really “green” (increased dependence on synthetic fertilizers and irrigation to increase crop yields per hectare at the expense of long-term soil health and yields per unit of input; increased dependence of indebted farmers on multinational producers of seeds, increasingly genetically modified, and which have not always delivered the touted benefits). It was thought that the sea was so immense that one needed not to worry about fishery limits, but now it is known that it is not always the case with many examples of overfishing of some populations. The 1980s were the time of the “Blue Revolution” of aquaculture development at sea, but it is also known that it is not always “green.” It is, consequently, time to make the “Blue Revolution” greener; it is time for the “Turquoise Revolution” to move aquaculture to a new ERA of Ecosystem Responsible Aquaculture at sea and on land, in seawater and freshwater, and in temperate and tropical regions.

### Acknowledgments

I greatly appreciate the support this work received from the Natural Sciences and Engineering Research Council of Canada (NSERC) strategic Canadian Integrated Multi-Trophic Aquaculture Network (CIMTAN) in collaboration with its partners, Fisheries and Oceans Canada, the University of New Brunswick, Cooke Aquaculture Inc., Kyuquot SEAfoods Ltd. and Marine Harvest Canada Ltd.

## Bibliography

### Primary Literature

1. FAO (2009) The State of World Fisheries and Aquaculture 2008. FAO of the United Nations, Rome, xvi + 176 p
2. Tacon AGJ, Metian M, Turchini GM, De Silva SS (2010) Responsible aquaculture and trophic level implications to global fish supply. *Rev Fish Sci* 18(1):94–105
3. Troell M, Halling C, Neori A, Chopin T, Buschmann AH, Kautsky N, Yarish C (2003) Integrated mariculture: asking the right questions. *Aquaculture* 226:69–90
4. Chopin T (2006) Integrated multi-trophic aquaculture. What it is and why you should care ... and don't confuse it with polyculture. *North Aquacult* 12(4):4
5. Edwards P (1993) Environmental issues in integrated agriculture-aquaculture and wastewater-fed fish culture systems. In: Pullin RSV, Rosenthal H, Maclean JL (eds) Environment and aquaculture in developing countries. ICLARM Conference Proceedings, vol 31, pp 139–170
6. Hambrey J, Edwards P, Belton B (2008) An ecosystem approach to freshwater aquaculture: a global review. In: Soto D, Aguilar-Manjarrez J, Hishamunda N (eds) Building an ecosystem approach to aquaculture. FAO/Universitat de les Illes Balears Experts Workshop. 7–11 May 2007, Palma de Mallorca Spain. FAO Fisheries and Aquaculture Proceedings 14. FAO, Rome, pp 117–221
7. Troell M (2009) Integrated marine and brackishwater aquaculture in tropical regions: research, implementation and prospects. In: Soto D (ed) Integrated mariculture: a global review. FAO Fisheries and Aquaculture Technical Paper 529. FAO, Rome, pp 47–131
8. Greenberg P (2010) Four fish – the future of the last wild food. Penguin Press, New York, 285 p
9. Timmons MB, Ebeling JM, Wheaton FW, Summerfelt ST, Vinci BJ (2002) Recirculating aquaculture systems (2nd edn). NRAC Publication No. 01-002. Cayuga Aqua Ventures, Ithaca, 769 p
10. Ayer NW, Tyedmers PH (2009) Assessing alternative aquaculture technologies: life cycle assessment of salmon culture systems in Canada. *J Cleaner Prod* 17:362–373
11. Skår CK, Mortensen S (2007) Fate of infectious salmon anaemia virus (ISAV) in experimentally challenged blue mussels *Mytilus edulis*. *Dis Aquat Org* 74:1–6
12. Pang SJ, Xiao T, Bao Y (2006) Dynamic changes of total bacteria and *Vibrio* in an integrated seaweed-abalone culture system. *Aquaculture* 252:289–297
13. Pang SJ, Xiao T, Shan TF, Wang ZF, Gao SQ (2006) Evidences of the intertidal red alga *Grateloupia turuturu* in turning *Vibrio parahaemolyticus* into non-culturable state in the presence of light. *Aquaculture* 260:369–374
14. Molloy SD, Pietrak MR, Bouchard DA, Bricknell I (2011) Ingestion of *Lepeophtheirus salmonis* by the blue mussel *Mytilus edulis*. *Aquaculture* 311:61–64

15. Johansson DJA, Azar CA (2007) A scenario based analysis of land competition between food and bioenergy production in the US. *Clim Change* 82:267–291
16. Naylor RL, Liska AJ, Burke MB, Falcon WP, Gaskell JC, Rozelle SD et al (2007) The ripple effect: biofuels, food scarcity and the environment. *Environment* 49:30–43
17. SOFA (2008) The state of food and agriculture. Biofuels prospects, risks and opportunities. FAO of the United Nations, Rome
18. Jordan N, Boody G, Broussard W, Glover JD, Keeney D, McCown BH, McIsaac G, Muller M, Murray H, Neal J, Pansing C, Turner RE, Warner K, Wyse D (2007) Sustainable development of the agricultural bio-economy. *Science* 316:1570–1571
19. Sawyer D (2008) Climate change, biofuels and eco-social impacts in the Brazilian Amazon and Cerrado. *Phil Trans R Soc B* 363:1747–1752
20. Koh LP (2007) Can palm oil plantations be made more hospitable for forest butterflies and birds? *J Appl Ecol* 44:703–713
21. Fargione J, Hill J, Tilman D, Polasky S, Hawthorne P (2008) Land clearing and the biofuel carbon debt. *Science* 319:1235–1238
22. Whitmarsh DJ, Cook EJ, Black KD (2006) Searching for sustainability in aquaculture: an investigation into the economic prospects for an integrated salmon-mussel production system. *Mar Policy* 30:293–298
23. Ridler N, Wowchuk M, Robinson B, Barrington K, Chopin T, Robinson S, Page F, Reid G, Szemerda M, Sewuster J, Boyne-Travis S (2007) Integrated multi-trophic aquaculture (IMTA): a potential strategic choice for farmers. *Aquac Econ Manage* 11:99–110
24. Costanza R, d'Arge R, de Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill RV, Paruelo J, Raskin RG, Sutton P, van den Belt M (1997) The value of the world's ecosystem services and natural capital. *Nature* 387:253–260
25. Feely RA, Sabine CL, Lee K, Berelson W, Kleypas J, Fabry VJ, Miller FJ (2004) Impact of anthropogenic CO<sub>2</sub> on the CaCO<sub>3</sub> system in the oceans. *Science* 305(5682):362–366
26. Thomsen B (2006) Freshwater aquaculture in Denmark: a new platform for sustainable growth. *Proceeding of the Canadian Freshwater Aquaculture Symposium, Aquaculture Canada 2004. Aquaculture Association of Canada Special Publication vol 11, pp 26–30*
27. Troell M, Halling C, Nilsson A, Buschmann AH, Kautsky N, Kautsky L (1997) Integrated marine cultivation of *Gracilaria chilensis* (Gracilariales, Rhodophyta) and salmon cages for reduced environmental impact and increased economic output. *Aquaculture* 156:45–61
28. Chopin T, Buschmann AH, Halling C, Troell M, Kautsky N, Neori A, Kraemer GP, Zertuche-Gonzalez JA, Yarish C, Neefus C (2001) Integrating seaweeds into marine aquaculture systems: a key towards sustainability. *J Phycol* 37:975–986
29. Lindhal O, Kolberg S (2009) Can the EU agri-environmental aid program be extended into the coastal zone to combat eutrophication? *Hydrobiologia* 629:59–64
30. Lindahl O, Hart R, Hernroth B, Kollberg S, Loo L-O, Olrog L, Rehnstam-Holm A-S, Svensson J, Svensson S, Syversen U (2005) Improving marine water quality by mussel farming: a profitable solution for Swedish society. *Ambio* 34:131–138
31. Ferreira JG, Hawkins AJS, Bricker SB (2007) Management of productivity, environmental effects and profitability of shellfish aquaculture – the Farm Aquaculture Resource Management (FARM) model. *Aquaculture* 264:160–174
32. Ferreira JG, Sequeira A, Hawkins AJS, Newton A, Nickell TD, Pastres R, Forte J, Bodoy A, Bricker SB (2009) Analysis of coastal and offshore aquaculture: application of the FARM model to multiple systems and shellfish species. *Aquaculture* 289:32–41
33. Gren IM, Lindhal O, Lindqvist M (2009) Values of mussel farming for combating eutrophication: an application to the Baltic Sea. *Ecol Eng* 35:935–945
34. Lackner KS (2003) A guide to CO<sub>2</sub> sequestration. *Science* 300:1677–1678
35. Nellemann C, Corcoran E, Duarte CM, Valdés L, De Young C, Fonseca L, Grimsditch G (2009) Blue carbon. A rapid response assessment. United Nations Environment Program, GRID-Arendal. [www.grida.no](http://www.grida.no)
36. Chung IK, Beardall J, Mehta S, Sahoo D, Stojkovic S (2010) Using marine macroalgae for carbon sequestration: a critical appraisal. *J Appl Phycol*. doi:10.1007/s10811-010-9604-9
37. Beardall J, Raven JA (2004) The potential effects of global climate change on microalgal photosynthesis, growth and ecology. *Phycologia* 43:26–40
38. Liu F, Pang SJ, Chopin T, Xu N, Shan TF, Gao SQ, Sun S (2009) The dominant *Ulva* strain of the 2008 green algal bloom in the Yellow Sea was not detected in the coastal waters of Qingdao in the following winter. *J Appl Phycol*. doi:10.1007/s10811-009-9489-7
39. Pang SJ, Liu F, Shan TF, Xu N, Zhang ZH, Gao SQ, Chopin T, Sun S (2010) Tracking the algal origin of the *Ulva* bloom in the Yellow Sea by a combination of molecular, morphological and physiological analyses. *Mar Env Res* 69:207–215
40. Neori A, Chopin T, Troell M, Buschmann AH, Kraemer GP, Halling C, Shpigel M, Yarish C (2004) Integrated aquaculture: rationale, evolution and state of the art emphasizing seaweed biofiltration in modern mariculture. *Aquaculture* 231:361–391
41. Chopin T, Robinson SMC, Troell M, Neori A, Buschmann AH, Fang J (2008) Multitrophic integration for sustainable marine aquaculture. In: Jørgensen SE, Fath BD (eds) *The encyclopedia of ecology, ecological engineering*, vol 3. Elsevier, Oxford, pp 2463–2475
42. Barrington K, Chopin T, Robinson S (2009) Integrated multi-trophic aquaculture (IMTA) in marine temperate waters. In: Soto D (ed) *Integrated mariculture: a global review*. FAO Fisheries and Aquaculture Technical Paper 529. FAO, Rome, pp 7–46
43. Brzeski V, Newkirk G (1997) Integrated coastal food production systems – a review of current literature. *Ocean Coast Manage* 34:55–71

44. Barrington K, Ridler N, Chopin T, Robinson S, Robinson B (2010) Social aspects of the sustainability of integrated multi-trophic aquaculture. *Aquac Int* 18:201–211
45. Shuve H, Caines E, Ridler N, Chopin T, Reid GK, Sawhney M, Lamontagne J, Szemerda M, Marvin R, Powell F, Robinson S, Boyne-Travis S (2009) Survey finds consumers support Integrated Multi-Trophic Aquaculture. *Effective marketing concept key*. *Global Aquaculture Advocate* 12(2):22–23
46. Chopin T, Sawhney M (2009) Seaweeds and their mariculture. In: Steele JH, Thorpe SA, Turekian KK (eds) *The encyclopedia of ocean sciences*. Elsevier, Oxford, pp 4477–4487
47. FAO (2010) *The State of World Fisheries and Aquaculture 2010*. FAO of the United Nations, Rome, xv + 197 p
48. Chopin T (1995) Mixed, integrated, poly-, or multi-level aquaculture – whatever you call it, it is time to put seaweeds around your cages! In: *Proceedings of the Conference on cold water aquaculture to the Year 2000*, St. Andrews, Canada, 6–8 September 1995. *Aquaculture Association of Canada, Special Publication No. 2* (1997):110
49. Chopin T (2008) Integrated multi-trophic aquaculture (IMTA) will also have its place when aquaculture moves to the open ocean. *Fish Farmer* 31(2):40–41
50. Troell M, Joyce A, Chopin T, Neori A, Buschmann AH, Fang JG (2009) Ecological engineering in aquaculture – potential for integrated multi-trophic aquaculture (IMTA) in marine off-shore systems. *Aquaculture* 297:1–9
51. Lee JH, O'Keefe JH, Lavie CJ, Harris WS (2009) Omega-3 fatty acids: cardiovascular benefits, sources and sustainability. *Nat Rev Cardiol* 6:753–758
52. Chopin T (2009) Let's not move just the fish to the open ocean. . . an integrated multi-trophic aquaculture approach should not be an afterthought for 2050. In: *The ecology of marine wind farms: Perspectives on impact mitigation, siting, and future uses*. 8th Annual Ronald C. Baird Sea Grant Science Symposium. 2–4 November 2009, Newport, USA. [http://seagrant.gso.uri.edu/baird/2009\\_windfarms/abstracts/chopin.pdf](http://seagrant.gso.uri.edu/baird/2009_windfarms/abstracts/chopin.pdf)
53. Subhadra B, Grinson G (2010) Algal biorefinery-based industry: an approach to address fuel and food insecurity for a carbon-smart world. *J Sci Food Agric*. doi:10.1002/jsfa.4207

### Books, Reviews, and Websites

- Costa-Pierce BA (2002) *Ecological aquaculture: the evolution of the blue revolution*. Blackwell Science, Oxford, 382 p
- Soto D, Aguilar-Manjarrez J, Hishamunda N (2008) Building an ecosystem approach to aquaculture. FAO/Universitat de les Illes Balears Experts Workshop. 7–11 May 2007, Palma de Mallorca, Spain. *FAO Fisheries and Aquaculture Proceedings* 14. FAO, Rome, 221 p
- Web site of Thierry Chopin: [www.unbsj.ca/sase/biology/chopinlab/](http://www.unbsj.ca/sase/biology/chopinlab/)
- Web site on IMTA on Wikipedia: [http://en.wikipedia.org/wiki/Integrated\\_Multi-trophic\\_Aquaculture](http://en.wikipedia.org/wiki/Integrated_Multi-trophic_Aquaculture)

## Aquaculture, Sustainability Science in

BARRY A. COSTA-PIERCE<sup>1</sup>, GLENN G. PAGE<sup>2</sup>

<sup>1</sup>Graduate School of Oceanography, Rhode Island Sea, Grant College Program, University of Rhode Island, URI Bay Campus, Narragansett, RI, USA

<sup>2</sup>SustainaMetrix, LLC, Johns Hopkins University, Baltimore, MD, USA

### Article Outline

Glossary  
 Definition  
 Introduction  
 Sustainability Strategic and Implementation Planning for Aquaculture  
 Improved Governance of Aquaculture Ecosystems  
 Social Ecology of Aquaculture  
 Future Directions: Sustainability Science Opportunities for Aquaculture  
 Determination of Sustainable Aquaculture for Retailers  
 Aquaculture and the Restoration of Ecosystems  
 Aquaculture and Agriculture Science  
 Aquaculture for the Poor  
 Bibliography

### Glossary

**Stewardship** Ecosystem stewardship is an ethic practiced by aquaculture practitioners, organizations, communities, and societies who strive to sustain the qualities of healthy and resilient ecosystems and their associated human communities. Stewardship takes the long-term view and promotes activities that provide for the well-being of both the present and future generations.

**Nested systems of governance** Environmental and societal issues relating to sustainable aquaculture impact, and are influenced by, conditions and actions (at both higher and lower levels) in an ecosystem governance hierarchy. Some issues can be addressed more effectively at one level, and less effectively at another. The choice of the issue or set of issues to be addressed within a given site must

therefore be made in full knowledge of how responsibility and decision-making authority are distributed within a layered governance system. Planning and decision-making for aquaculture at one scale, for example, within a municipality or province, should not contradict or conflict with planning and management at another scale, for example, planning for large-scale aquaculture at the nation-state scale. The reality is that such contradictions and conflicts are common. A major challenge for the aquaculture practitioner is to recognize these differences and work to either change them or select goals and strategies that recognize that such contradictions must be accommodated or resolved. In practical terms, this means that a central feature of ecosystem-based aquaculture is that all planning and decision-making must recognize and analyze conditions, issues, and goals in respect to the next higher level in a governance system. Thus, ecosystem-based aquaculture at the municipal scale must – at a minimum – be placed within the context of governance at the scale of the province.

**Participation** One of the defining characteristics of the practice of the ecosystem approach to aquaculture is its emphasis on participation and its relevance to the people affected. The emphasis upon participation recognizes that if an aquaculture program is to be successful, those whose collaboration and support is needed must be involved in the processes of defining the issues that the program will address, and in selecting the means by which goals and objectives will be achieved. Both individuals and members of communities and institutions are more likely to comply with a management program when they feel that it is consistent with their values, responds to their needs, and to their beliefs of how human society should function. Voluntary compliance by a supportive population lies at the heart of the successful implementation of a program. A participatory approach helps stakeholders and the public to see the efforts of an aquaculture program as a whole.

**Area of focus** The area of focus (AoF) is the geographically defined area that an ecosystem-based aquaculture project or program has decided to address and that therefore is the focal point for a baseline.

The term “area of focus” is a geographic limit set to model the choices available to the aquaculture practitioner and allows for a dialogue between stakeholders as to the influence of the production. The AoF is a simplification of the far more complex concept of an “action arena” put forward by Ostrom [1] to model the choices of individuals when studying the behavior of institutions.

**Adaptive management** A central feature of the practice of any form of ecosystem-based aquaculture is that it must respond positively to changing conditions within its AoF (and to its own experience). In other words, the practice of aquaculture must be grounded in a process of learning and adaptation (the “evolution of the blue revolution” [2]). Adaptive management is not reactive management, but proactive thinking and acting. This does mean that the aquaculture practitioner simply responds to the unexpected. Adaptive management in aquaculture is a conscious process of examining the course of events as these events are revealed by preselected indicators of changes in an aquaculture ecosystem (both its social and environmental components), and by events occurring at differing spatial scales.

**Capacity building** There is growing international recognition that the lack of human capacity to practice an ecosystem approach to aquaculture is a key factor in limiting forward progress in the conservation and sustainable use of aquatic systems [3, 4]. To date, however, no accepted performance standards have been developed for assessing the effectiveness and impacts of aquaculture projects and programs that have adopted the ecosystem approach. Conceptual frameworks and methods for assessing the maturity of aquaculture development and management initiatives, and gauging their impacts upon the condition of coastal ecosystems are offered herein. These are the core ingredients for an ecosystem’s approach to aquaculture that builds the capacity of local populations and leaders to identify forces that shape the coastal ecosystems of which they are a part, and to select the actions that can maintain and enhance qualities that are critical to a desirable future.

**Carrying capacity** The carrying capacity is the number of organisms or farming operations that the



environment can sustain indefinitely without environmental harm, given the food, habitat, space, water, and other requirements from the environment.

**Precautionary principle** A principle that states that if an action or policy has a suspected risk of causing harm to the public or to the environment that in the absence of scientific consensus the burden of proof rests on those who advocate taking the action.

**Sustainable development** The management and conservation of the natural resource base and the orientation of technological and institutional change in such a manner as to ensure the attainment and continued satisfaction of human needs for present and future generations. Sustainable development conserves resources, is environmentally non-degrading, and is technically appropriate, economically viable, and socially acceptable [5].

**Transdisciplinary** A modern research strategy that crosses many disciplinary boundaries to create a holistic approach. Transdisciplinary research efforts are focused on problems that cross the boundaries of two or more disciplines, and develop new or reframe old concepts, methods, and findings that were originally developed by one discipline, but are now used by several others.

## Definition

There is no one definition of “sustainability” as the concept applies to aquaculture. Most aquaculture scientists define sustainability as synonymous with “environmental sustainability.” Sustainable aquaculture is however a concept broader than determinations of site-specific environmental impacts since it embodies a scientific knowledge of systematic impacts of aquaculture off-site, and impacts to combined human-environmental systems. Sustainable aquaculture incorporates the concepts of “stewardship,” “design with nature,” the “precautionary principle,” “risk analysis,” and “carrying capacity.” Sustainability science in aquaculture is used to undertake more comprehensive planning for multiple impacts on multiple time and spatial scales to better understand and plan for the consequences of aquaculture development options.

## Introduction

- “The changes taking place [on planet Earth] are, in fact, changes in the human-nature relationship. They are recent, they are profound, and many are accelerating. They are cascading through the Earth’s environment in ways that are difficult to understand and often impossible to predict. Surprises abound” [6].

There are many definitions of “sustainability” as the concept applies to aquaculture. The most popular definition of sustainable development is to “meet present needs without compromising the ability of future generations to meet their needs” adopted at a United Nations conference in 1987. Most definitions of sustainability are synonymous with “environmental sustainability” of air, water, and land systems. Sustainability is however a concept broader than examining the site-specific environmental impacts of externalities in planning for site-specific developments; it also accounts for systematic impacts off site, and impacts to combined human-environmental systems for food, water, waste, energy, and shelter. The many definitions of sustainability all embody in common, the concepts of “stewardship,” “design with nature,” plus incorporate recent concepts of the “precautionary principle,” and “carrying capacity.” Sustainability science uses the wisdom from multiple disciplines in decision-making (e.g., it is “transdisciplinary”). In aquaculture, it is used to undertake more comprehensive planning for multiple impacts on multiple time and spatial scales to better understand and plan for the consequences of development options.

The emerging fields of ecological aquaculture [2,3] and agroecology [7, 8] recognize that the implementation of more sustainable food production systems require knowledge about how ecosystems are utilized and how conflicts among social groups are addressed. A baseline of response to social-ecological changes is the foundation for the implementation of more sustainable food systems, and the practice of adaptive management must be included as responses to changes in the condition of ecosystems in which new food production is conducted requires incorporation of an iterative learning process.

The use of sustainability science in aquaculture marks the path toward encouraging a long-term

perspective and an appreciation of the roles played not only by ecologists, but also by civil societies, markets, and governments in adapting to food systems and ecosystems changes. The use of sustainability science in aquaculture is an approach that is fundamentally a knowledge-based enterprise that incorporates baseline information on natural and human ecosystems, then develops, evaluates, encourages, and communicates imagination, ingenuity, and innovation at both the individual and institutional levels [9].

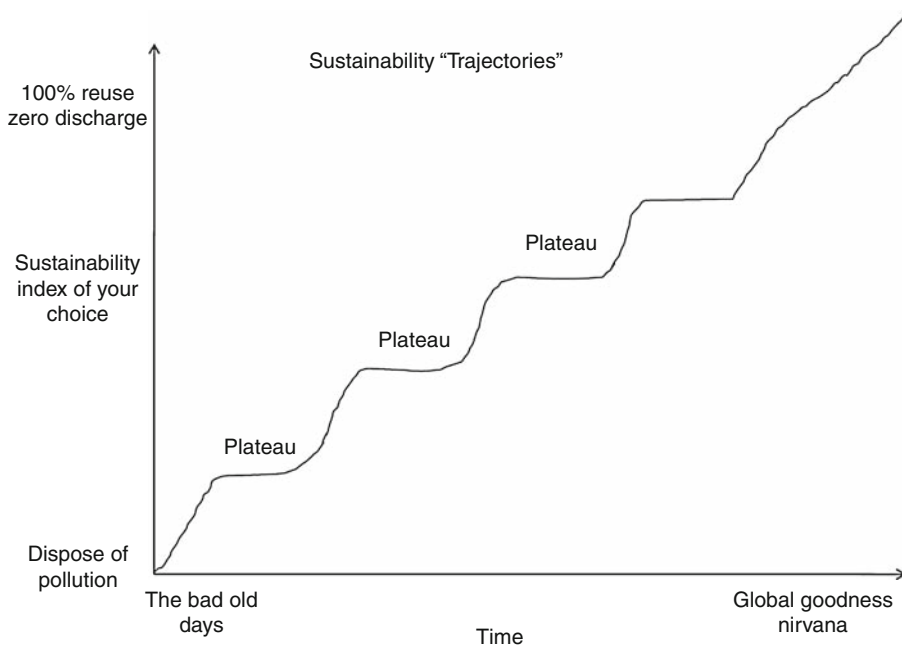
This information is designed for use by teams of aquaculture professionals working to apply the principles of ecosystem-based management. Information obtained is typically cross sectoral as interdisciplinary groups are needed that are educated in such diverse fields as the natural and social sciences, law, and business. Applying the notions of sustainability science in aquaculture is intended to inspire engagement of governmental agencies, businesses, nongovernmental groups, and academics to achieve the highest form of sustainable development in any known protein

production food system by using the concepts of ecological design and through the many forms of stewardship. At present, there is a paucity of information targeted specifically for those engaged in aquaculture programs and projects in places where the ability of government to regulate and direct the processes of ecosystem change is weak or severely constrained.

### Sustainability Strategic and Implementation Planning for Aquaculture

The concept of sustainability and the methods to measure the evolutionary progress toward more sustainable systems are limited, but have become a necessity. Wurts [10] stated that “Whether the word sustainability has become overused or not, it has catalyzed a forum for oversight of the growth and development of aquaculture on a global scale.”

Sustainability is not a “black/white” phenomena; rather, it is many “shades of gray,” an evolutionary process that is called the “sustainability trajectory” (Fig. 1). To measure and evaluate progress along



**Aquaculture, Sustainability Science in. Figure 1**

Sustainability is neither a “black or white,” nor an “either or” concept. It is the evolution of practices and principles over time toward ameliorating environmental and social impacts, with plateaus along the way in changed states. In many cases, these “pauses” are done to ensure economic viability. In this diagram, one example indicator (water use, other important indicators have been proposed [17]) is plotted along such a “sustainability trajectory”

a trajectory requires establishment of baselines for the main issues of public concerns, then developing a diverse but targeted set of resource and social indicators. These indicators are then used to report progress on and analyze interactions between social, environmental, and economic impacts (both positive and negative ones). It is important to note that sustainability science as applied to aquaculture is driven as much by social as by environmental/ecological concerns; thus, sole involvement of technical experts in sustainability plans and assessments is insufficient.

Developing an operational framework for how the sustainability of aquaculture operations operates is the first step. Having such a blueprint is rare for aquaculture businesses and management entities, and is very much needed. There are numerous certification bodies that are vying for the opportunity to use their labels/logos to claim ownership of the sustainability rubric in aquaculture. An overall sustainability science approach is proposed, which can step above the cacophony of approaches and assist in developing a common language and can be used by international and national, non-advocacy organizations such as the FAO, ICES, or governments and industry.

The approach used here is based upon the development of a baseline that has two parts and then follows a sequence of five steps:

The *first part of a baseline* is an ecosystem audit of the AoF that defines the natural and social systems within which aquaculture is planned.

This involves the documentation and analysis of both natural and social systems, draws upon case studies of other aquaculture systems in the region and how the governance system in a specific place has responded – or failed to respond – to the trajectories of ecosystem change. It examines the long-term trends in both human well-being and the environmental conditions in the AoF and examines responses to the issues raised by past and current expressions of food production systems.

The *second part of the baseline* is an outline of the strategic approach to designing a new aquaculture program, or adapting an ongoing program, to address the ecosystem management issues of the place in terms of economic, environmental, and societal benefits. Together, these parts form the reference points against which future changes in the aquaculture ecosystem will

be gauged. These methods encourage a long-term perspective, an appreciation of the roles played by civil society, markets, and government, and offer a holistic, ecosystem-based, approach to stewardship.

Baselines are not formulaic but are designed planning exercises with buy-in from key stakeholders such as the client, community, regulatory community, or identified group of people involved in the project. While not formulaic, baselines do comprise a set of common metrics to include

- Ecological aquaculture design (or redesign) of production practices (see Ecological Aquaculture chapter in this Encyclopedia)
- Health and quality control standards
- Social goals at both the individual and community levels for local food, job, and regional development (e.g., “green jobs,” “local foods”)
- Governance goals

The following five steps encompass some essential parts of any baselining process:

1. *Define the sustainability issues.* Aquaculture systems can use environmentally derived feeds, water, and energy, occupy land and water space, and generate wastes. There are at least eight issues of wide public and regulatory concerns regarding aquaculture development:
  - Destruction of habitats
  - No net gain to global seafood supplies
  - Environmental impacts of discharged wastes
  - Impacts of escapees
  - Diseases in farmed fish
  - Chemical use and discharge
  - Impacts on coastal marine mammals
  - Improper siting causes visual pollution

Once issues are defined, a baseline can be further developed which can measure progress over time by
2. *Completing a sustainability assessment* of these issues by evaluating the status of current aquaculture practices that affect natural and social resource systems (Table 1), which also includes an assessment of governance systems (Tables 2 and 3) [11–13]
3. *Completing a detailed risk analysis* for all components of this comprehensive assessment [14]

**Aquaculture, Sustainability Science in. Table 1** Important sustainability science tools used to assess and communicate the sustainability of aquaculture<sup>1</sup>

| Social Sustainability                                                                                                                                                                                               | Environmental Sustainability                                                                                                                                                                                                                                                                                                                    | Economic Sustainability                                                                                                                                                                                                                                                                                                                          |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Stakeholder analysis:</b> analysis of attitudes of stakeholders at the initiation of and throughout a project. Allows tracking of how stakeholders change attitudes over time with educational processes [46–50] | <b>Life cycle analysis:</b> complete assessment of products from raw material production, manufacture, distribution, use and disposal, including all transportation; used to optimize environmental performance of a single product or a company. A similar analysis called a MET (Materials, Energy, and Toxicity) Matrix is also used [53–55] | <b>Cost-benefit analysis:</b> analysis of cost effectiveness of different uses to determine if benefits can outweigh costs [59]                                                                                                                                                                                                                  |
| <b>ISO 26000 guidelines</b> for corporate social responsibility [51]                                                                                                                                                | <b>ISO 14000 certification:</b> norms to promote more effective and efficient environmental management and provide tools for gathering, interpreting, and communicating environmental information [56]                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                  |
| <b>ICLEI</b> (International Council for Local Environmental Initiatives) provides software and tools to help local governments achieve sustainability goals [52]                                                    | <b>Environmental impact assessment:</b> the process of identifying, predicting, evaluating, mitigating biophysical, social, and other effects of development proposals prior to policy decisions [57, 58]                                                                                                                                       | <b>Triple bottom line or “full cost” accounting:</b> costs considered for all environmental, economic, and social impacts; costs measured in terms of opportunity costs (the value of their best alternative use); guiding principle is to list all parties affected and place a monetary value on effects on welfare as valued by them [60, 61] |
|                                                                                                                                                                                                                     | <b>Environmental indicators:</b> the use of quantitative indicators of resource use, efficiency, and waste production in aquaculture [17]                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                  |

<sup>1</sup>This table does not contain a comprehensive list of all available tools; rather, tools selected here were chosen since they appear regularly in the modern sustainable aquaculture research, industry, and management literature. Gibson et al. [62] give a most complete analysis of all of the available tools for sustainability assessments.

4. *Completing a plan for ameliorating identified impacts* by incorporation of better (or best) practices [15–17], and/or enhancing reuse or recycling pathways, and
5. *Completing a plan for communicating* the evolution of operations toward greater stewardship and sustainability [14]

To be effective, sustainable aquaculture initiatives must: (a) be “profitable” over long periods of time – ideally many decades; (b) be capable of being adapted to changing conditions; and (c) provide the mechanisms to encourage both wise resource use and collaborative behaviors. Much of the challenge

lies in achieving changes in the behavior of those who may be unaware of the benefits of sustainable aquaculture.

Sustainable aquaculture integrates the best available science with a transparent, equitable, and democratic approach to planning and decision-making. This ecosystem approach to management needs to be carried out in a strategic manner that tailors principles of good practice to the culture and the needs of a specific place. Successful, sustainable aquaculture operations advance through linked cycles of planning, implementation, and reassessment. These features of ecosystem management signal the transition from traditional sector-by-sector planning and decision-making to a more holistic

**Aquaculture, Sustainability Science in.** Table 2 Sustainability science assessments of aquaculture includes an assessment of governance systems, which examine the three processes of governance: the marketplace, the government, and civil society

| Major Expressions of Governance                      |
|------------------------------------------------------|
| <i>Government</i>                                    |
| Laws and regulations                                 |
| Taxation and spending policies                       |
| Education and outreach                               |
| <i>Marketplace</i>                                   |
| Profit seeking                                       |
| Ecosystem service valuation                          |
| Cost-benefit analysis                                |
| Ecolabeling and Green Products                       |
| <i>Civil Society: Organizations and Institutions</i> |
| Product choices                                      |
| Advocacy and lobbying                                |
| Vote casting                                         |
| Comanagement                                         |
| Stewardship activities                               |

approach based on the interactions between sectors and within and among ecosystems.

Aquaculture that is constructed upon principles that encourages high-energy consumption and the profligate use of natural resources must give way to new locally derived values and new forms of practice. As suggested by Daly [18], qualitative development rather than quantitative growth is the path of future progress. If such ideas are to be made operational at the scale of an aquaculture operation, a trajectory can be established based on goals for profit as well as social and environmental benefit. Once the goals of an aquaculture program or project have been defined as expressions of the ecosystem approach, much of the day-to-day work is concerned with the well-known best practices of aquaculture management.

For example, there has been much debate about the impacts of shrimp pond mariculture on mangrove forests through the Topics. Mangrove ecosystems

provide essential goods and services to humanity, harboring an extraordinarily large biodiversity for the small areas of the planet that these systems occupy, and provide a sustainable source of timber and charcoal to coastal communities while protecting fragile coastlines from erosion and storms. Establishment of proper scientific baselines to measure the true impacts of mariculture on coastal ecosystems is essential. Pullin [19] cautions that, “Analysis on depletion of mangrove cover in Asia point towards the fact that shrimp ponds have recently been and/or now being constructed either on former mangrove areas that were cleared long ago and considered degraded), or on more recently cleared areas for which the primary purpose of clearance was timber abstraction (logging, wood chip industries or charcoal production) or by adopting traditional trapping ponds. . . . Aquaculturists in Asia are therefore more often than not the end users of already degraded or destroyed mangroves rather an the primary culprits of mangrove destruction.”

Good examples globally of an ecosystem approach to aquaculture at the watershed/aquaculture zone scale are found in both Israel and Australia. Both nations face severe land, water, and energy constraints. In Israel, highly efficient, landscape-sized integrations of reservoirs with aquaculture and agriculture have been developed [20, 21], as well as highly productive, land-based aquaculture ecosystems for marine species [22]. These aquaculture ecosystems are productive, semi-intensive enterprises that are water and land efficient, and are net energy and material gains to society. Aquaculture ecosystems are organized following well-established ecological principles similar to the fields of agroecology and agroecosystems [23].

In Australia, an Ecologically Sustainable Development (ESD) framework approach to aquaculture development was used [24]. This ESD framework identified important issues, developed comprehensive reports for each issue, and then prioritized each using risk assessments. The ESD process employed extensive community consultation that considered social and environmental values of all other marine users, and users’ management plans for operations and administration as well as environmental administrative attributes, then proposed development and monitoring plans.



**Aquaculture, Sustainability Science in. Table 3** Orders of governance outcomes [12, 13] applied to an ecosystem approach to aquaculture [2, 3]

| Orders       | Explanations                                                                                                                                                                                                                                       | Indicators                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| First Order  | Government at the national level commits to a plan of action designed to adopt an ecosystem approach to aquaculture (EAA) by issuing a formalized commitment to an EAA, thereby putting in place the “enabling conditions”                         | New laws, programs, and procedures are initiated that provide the legal, administrative, and management mechanisms to achieve the desired changes in behavior by:                                                                                                                                                                                                                                                                                                                              |
|              |                                                                                                                                                                                                                                                    | (i) Building constituencies that actively support EAA with the user groups that will be most affected; with government institutions involved; and with the general public                                                                                                                                                                                                                                                                                                                      |
|              |                                                                                                                                                                                                                                                    | (ii) Developing a formal government mandate for an EAA with the authority necessary to implement actions in the form of laws, decrees, or other high-level administrative decisions that create an EAA as a permanent feature of the governance structure of aquaculture; creation of commissions, working groups, user organizations, and nongovernmental organizations (NGOs) dedicated to the advancement of an EAA agenda; designating EAA zones                                           |
|              |                                                                                                                                                                                                                                                    | (iii) Devoting resources, especially sustained annual funding, adequate to implement an EAA                                                                                                                                                                                                                                                                                                                                                                                                    |
|              |                                                                                                                                                                                                                                                    | (iv) Developing an implementation plan of action for an EAA that is constructed around unambiguous goals                                                                                                                                                                                                                                                                                                                                                                                       |
|              |                                                                                                                                                                                                                                                    | (v) Creating the institutional capacity necessary to implement the new EAA plan of action                                                                                                                                                                                                                                                                                                                                                                                                      |
| Second Order | Evidence of successful implementation of an EAA                                                                                                                                                                                                    | 1. Changes in the behavior of institutions and interest groups have occurred such as collaborative planning and decision-making through creation of task forces, commissions, civic associations, etc.                                                                                                                                                                                                                                                                                         |
|              |                                                                                                                                                                                                                                                    | 2. Successful application of conflict mediation activities                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|              |                                                                                                                                                                                                                                                    | 3. Evidence of functional changes such as establishment of new public–private partnerships, new collaborative actions undertaken by user groups, implementation of new school curricula that incorporates an EAA                                                                                                                                                                                                                                                                               |
|              |                                                                                                                                                                                                                                                    | 4. Changes in behaviors directly affecting ecosystem goods and services, such as the elimination of socially and environmentally destructive aquaculture practices                                                                                                                                                                                                                                                                                                                             |
|              |                                                                                                                                                                                                                                                    | 5. Investments in infrastructure supportive of EAA policies and plans                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Third Order  | Evidence of sustained achievements in institutional and behavioral change due to an EAA as seen in the environment and indicators for the quality of life, incomes, or engagement in alternative livelihoods that have improved target communities | 1. Improvements in ecosystem qualities, such as sustained conservation of desired ecosystems and habitats, halting or slowing undesired trends such as nutrient releases, feed wastage, diseases, damaged benthic ecosystems, etc.                                                                                                                                                                                                                                                             |
|              |                                                                                                                                                                                                                                                    | 2. Improvements in society as evidenced by monitoring of social indicators such as increases in indices of quality of life, reduced poverty, greater life expectancy, better employment opportunities, greater equity in access to coastal resources, and the distribution of benefits from their use, greater order, transparency, and accountability in how planning and aquaculture development decision-making processes occur, greater food security, or greater confidence in the future |

As a result of this ESD approach, nine marine aquaculture zones of 2,400 ha in Port Phillip Bay and Westernport, Victoria, Australia were permitted. The Australian ESD approach combined analytical and participatory methods and developed sustainability plans that considered both ecosystem and human well-being, then developed implementation strategies by designing and enhancing effective governance systems for the expansion of aquaculture.

The development of a sustainability baseline should be the responsibility of a lead aquaculture agency. Its full implementation may require alternative methods of governance and employ innovative management approaches. There may be a need to facilitate an operational definition of aquaculture ecosystem boundaries for assessment, or area of focus, to set geographical limits to assess parameters such as carrying capacity or water management needs, and to understand the governance regime within which the area of focus is nested in order to understand and clarify such things as administrative and legal jurisdictions.

Using such guidance and sustainability science frameworks, the possibilities for designing productive aquaculture ecosystems that better fit into the local social and ecological context are many, since aquaculture can encompass the wide availability of species, environments, and cultures.

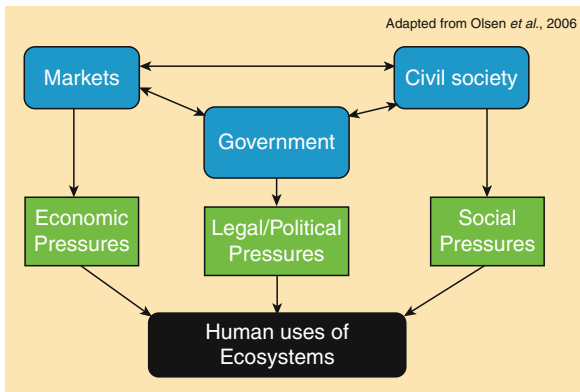
### **Improved Governance of Aquaculture Ecosystems**

To be effective, ecosystem-based aquaculture initiatives must (1) be sustainable over long periods of time – ideally over many decades, (2) be capable of being adaptable to changing conditions, and (3) provide the mechanisms to encourage or require specified forms of resource use and collaborative behaviors among institutions and user groups that are stakeholders of the aquaculture system. Much of the challenge lies in both understanding and achieving changes in the behavior of the stakeholder groups and institutions associated with the aquaculture production systems. Ecosystem-based aquaculture integrates the best available science with a transparent, equitable, and democratic approach to planning and decision-making. Management needs to be carried out in a strategic manner that tailors

principles of good aquaculture practice to the culture and the needs of a specific place. Successful aquaculture programs advance and change through linked cycles of planning, implementation, and reassessment. These features of ecosystem management signal the transition from traditional food production sector planning and decision-making to a holistic approach based on the interactions between sectors and within and among ecosystems.

FAO Fisheries and Aquaculture Department [25] found that one of the key trends toward more sustainable forms of aquaculture development and management is enhanced regulation and better governance. Governance is defined as the formal and informal arrangements, institutions, and mores that structure and influence how resources or an environment are utilized, how problems and opportunities are evaluated and analyzed, what behavior is deemed acceptable or forbidden, and what rules and sanctions are applied to affect how natural resources are distributed and used.

As shown in Table 2, there are three mechanisms by which the processes of governance are expressed: the marketplace, the government, and the institutions and arrangements of civil society [11]. These mechanisms interact with one another through complex and dynamic interrelationships that are examined and contrasted and documented in a baseline. Each of the three governance mechanisms influence and can alter patterns of behavior through measures such as those identified in Fig. 2. For sustainable, ecosystems-based aquaculture, it is important to distinguish between management and governance. Management is the process by which human and material resources are harnessed to achieve a known goal within a known institutional structure. Aquaculture business management, park management, personnel management, or disaster management is therefore spoken about. In these instances, the goals and the mechanisms of administration are well known and widely accepted. Governance, in contrast, addresses the values, policies, laws, and institutions by which a set of issues are addressed. It probes the fundamental goals and the institutional processes and structures that are the basis for planning and decision-making. Governance sets the stage within which management occurs [12].



**Aquaculture, Sustainability Science in. Figure 2**  
The three mechanisms by which the processes of governance are expressed interact with one another through complex and dynamic interrelationships that are vital parts of sustainability science assessments of aquaculture as each alter behaviors and decision-making that determine human uses of ecosystems [11–13]

The future of sustainable aquaculture is highly dependent on understanding the response by all three expressions of governance: markets, civil society, and government. For example, Kenya has fostered a participatory policy formulation for aquaculture, providing a legal and investment framework through government, establishing public–private partnerships to engage markets, providing basic infrastructure support, promoting self-regulation, providing a research platform for civil society to be engaged, undertaking zoning for aquaculture, and providing monitoring and evaluation support [25].

Adaptation of sustainability frameworks used to evaluate the needs and progress of governance on coastal management plans are essential to evaluate progress toward an ecosystem approach to aquaculture and build in adaptive learning and action into the strategic planning process. Governance frameworks recognize not only the importance of changes in practices such as changes over time in aquaculture farming ecosystems, but also recognize that for each change, there are correlated changes in the behavior of key partners and stakeholders within the sphere of influence of the management activity, and that these

changes can be measured at local, regional, and national levels (Fig. 3, Table 3).

Sectoral agencies responsible for managing activities impacting aquatic ecosystems (e.g., capture fisheries, coastal zone development, watershed management organizations, agriculture, forestry, and industrial developments) will have to develop new ways of interacting to regularly communicate, cooperate, and collaborate. The need for innovative governance to implement an ecosystem-based approach to aquaculture can be seen as an obstacle but can also be seen as an opportunity to increase the social benefits that are likely to develop through synergies among food production sectors.

### Social Ecology of Aquaculture

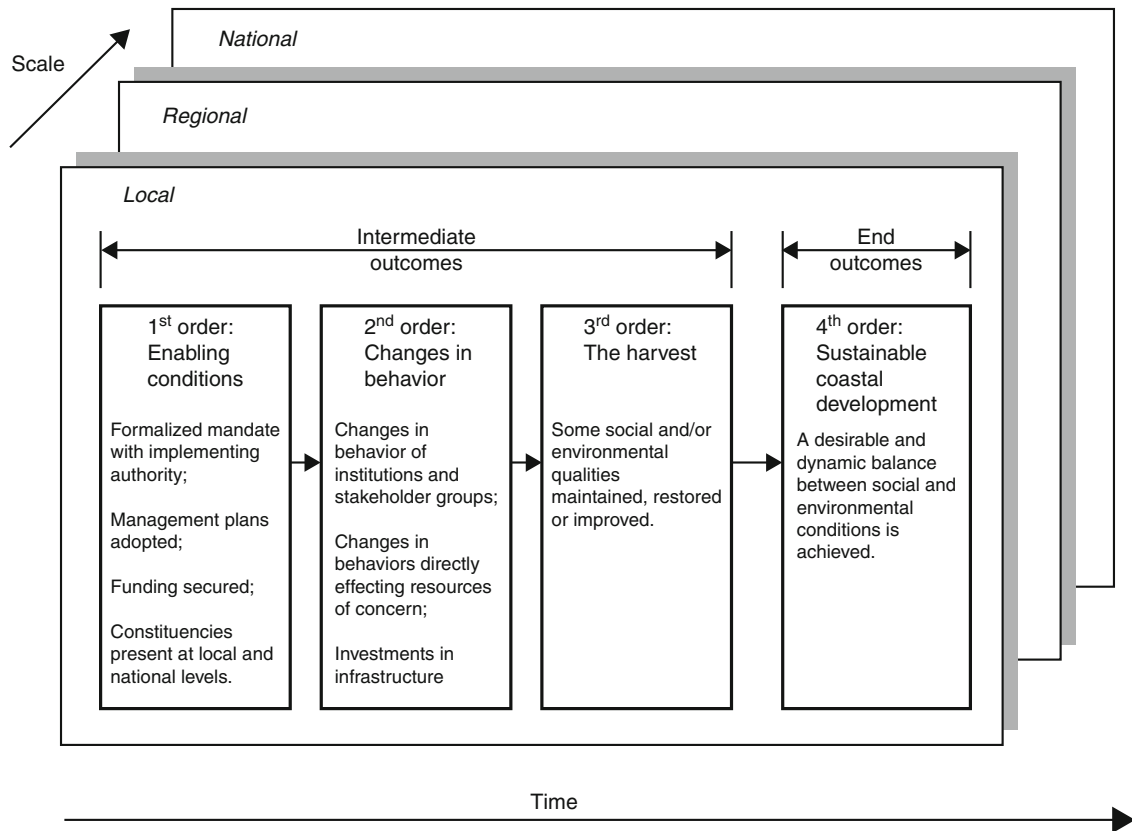
While there is much information on the natural ecology of food-producing ecosystems, there are few comprehensive frameworks for capturing the necessary social ecology of aquaculture.

Cadenasso et al. [26] have developed a “Human Ecosystem Framework” that could contribute to a baseline approach and assist in organizing multidisciplinary, social–ecological approaches to aquaculture development (Fig. 4). The most sustainable growth trajectories for aquaculture are to move toward more sustainable, social–ecological approaches to development; to shift patterns of production and consumption patterns from global to bioregional and local foods production and job creation; and to develop the indigenous human and institutional capacities that clearly demonstrate to society that “aquaculture is culture.”

### Future Directions: Sustainability Science Opportunities for Aquaculture

There are at least four major opportunities for sustainability science in the field of aquaculture in the

1. Determination of “sustainable aquaculture” for retail seafood companies
2. Growing fields of marine ecosystem and habitat restoration, conservation biology, and ecology
3. Accelerated use of agricultural meals and oils, and
4. Development of sustainable aquaculture for the poor



**Aquaculture, Sustainability Science in. Figure 3**

The four orders of coastal governance outcomes. This framework is used to develop governance baselines in environmental programs [12, 13]. An example of how progress toward better governance for sustainable aquaculture is shown in Table 3

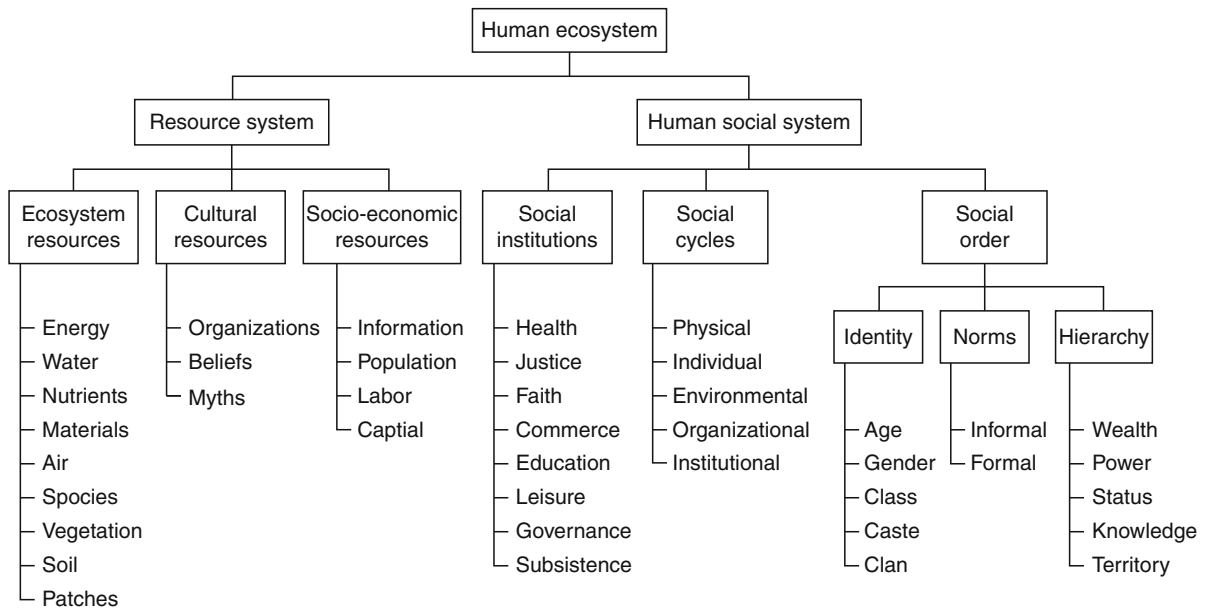
### Determination of Sustainable Aquaculture for Retailers

Sustainability science approaches to aquaculture can be used to better plan and develop aquaculture production networks for multiple species. Such planning approaches can be used to plan for the creation of highly diversified, segmented aquaculture networks, for maximal job creation at every unit step from “farm to plate” (e.g., seafood value chain planning), by creating numerous interconnections supplying inputs and outputs using local resources and recycled wastes and materials and expertise, and to close “leaky” loops of energy and materials that can potentially degrade natural ecosystems.

Behavioral changes will be required by industry. Social investments, strategic incentives/subsidies, and

innovative market mechanisms can help facilitate change in behaviors. Self-regulation by the aquaculture industry has led to codes of practice and better management practices.

Sustainability assessments are predicated upon the fact that the modern aquaculture industry desires to be seen as a responsible steward. This means going beyond “meeting the regulations.” There are a cacophony of certification bodies and seafood watch cards – there are an estimated 200 sustainable seafood guides available internationally – which has created a far too complex and sometimes conflicting recommendations to both consumers and retailers on what is “sustainable seafood” [27, 28]. Roheim [27] states that “Shrimp, in some form, appears as a green, yellow, red, and non-consensus list item” in the seafood “watch cards.”



**Aquaculture, Sustainability Science in. Figure 4**

The Human Ecosystem Framework [26]. Assessment of new interventions such as aquaculture into societies requires knowledge of not only biophysical and natural resource systems but also social resources and human social systems

The logic behind consumer approaches is that informed consumers who care about sustainable seafood will demand aquaculture products that carry a label or fit into the “green” (buy) area of a watch card, as opposed to those products that do not have the label, sending a market signal back to aquaculture industries that only products from sustainable aquaculture farms are preferred. Many of the independent certification programs that have developed ecolabels and “seafood watch cards” to provide consumers with additional information come from nongovernmental organizations (NGOs) with specific advocacy agendas and not from neutral, scientific sources, or from regulatory bodies charged with protecting the environment and society. For example, many fisheries and aquaculture scientists are deeply concerned that consumer recommendations of NGOs are moving demand (and use) from farmed stocks to already overburdened wild fisheries. The Monterey Bay Aquarium’s (MBA) Seafood Watch Program has produced millions of folding wallet cards featuring a “stop light” system of green (sustainable), yellow (choose with caution), and red (do not choose) recommendations. Farmed shrimp and salmon, two of the world’s largest aquaculture

industries are on the MBA red list. Roheim [27] mentions that the Compass Group a major food service company has used the MBA cards to decrease purchases of farmed shrimp and salmon, which, in effect, has created additional fishing pressure on wild shrimp and salmon stocks.

Most organizations believe that consumers’ increasing awareness of environmental and food safety issues will lead them to accept a wide variety of standards and labels, most of which are specifically intended to allay consumers’ concerns about negative environmental consequences.

However, Roheim [27–29] points out concerns over ecolabeling, especially the lack of transparency and opportunity for participation in the development of standards, and concerns of developing countries that ecolabeling schemes are an attempt at disguised protection of domestic industries to restrict market access and erode competitiveness. In addition, Wessels et al. [30] found that successful ecolabeling programs must accelerate consumer education programs so that consumers become more aware of differences in species, geographic regions, and certifying agencies.



Roheim [29] states that ecolabels require traceability. Traceability is the ability to follow the movement of a food through specified stages of production, processing, and distribution. Essentially, it is a record-keeping system that identifies and tracks products, transportation of products, and ingredients of products from origin to consumption, while providing the ability to quickly trace back products at any point along the supply chain. It is necessary for food safety purposes, in order to track backward in the food chain the source of food that made consumers ill, so that products could be removed from store shelves.

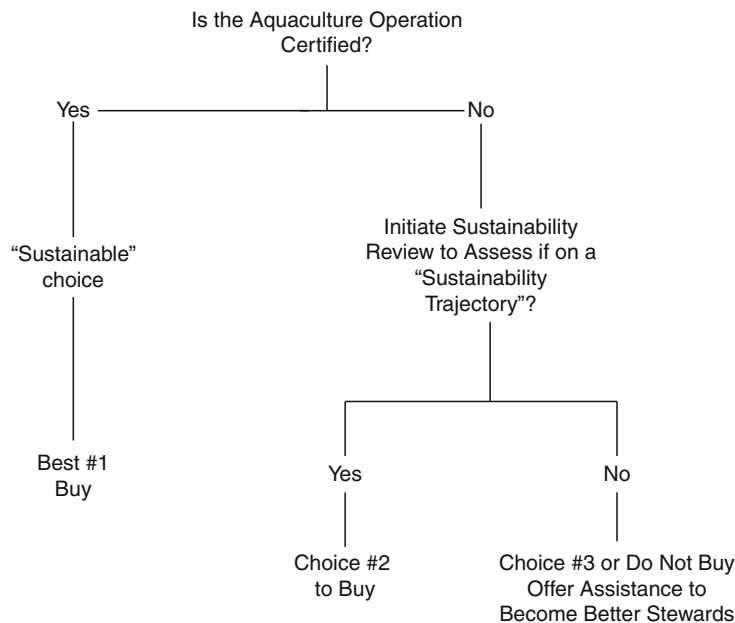
For consumers, Roheim [27] argues the need for ecolabels determined at the larger international levels, such as the Marine Stewardship Council (MSC), so that consumers “do not need to inquire about catch area or gear types, but only need to look for the label.” The plethora of efforts has also confused and perplexed retailers who are the main “drivers” of certification at present, not consumers [27–29]. However, even though many buyers wish to purchase sustainable seafood, most seafood products are not certified, and they are very confused by the many NGO efforts. A purchasing policy determined by assessing which seafoods are “sustainable” by making an assessment of the plethora

of NGOs, “opinions” seems confusing, risky, and costly. Rather, a simple, buying protocol (Fig. 5) that incorporates a sustainability assessment (where needed) as discussed here is recommended.

### Aquaculture and the Restoration of Ecosystems

Aquaculture science can be viewed as a “toolbox” with great potential for restoring aquatic ecosystems. There is an unbalanced focus on marine animal husbandry (e.g., “fed” aquaculture) causing a lack of appreciation for the positive environmental attributes of nonfood aquaculture such as marine agronomy, endangered species aquaculture, and aquaculture for environmental enhancement and rehabilitation, all of which use modern marine hatchery and nursery aquaculture practices [31].

Aquaculture technologies (hatchery, nursery, grow-out) for marine plants are used for the restoration of mangroves, sea grasses, and coastal wetland plants such as *Spartina sp.* In addition, live rock and coral aquaculture facilities are active for not only the aquarium trade, but also for the environmental restoration of coral reefs (liveaquaria.com). In this regard, there is little difference between sustainable aquaculture and the emerging



**Aquaculture, Sustainability Science in. Figure 5**

A simple decision tree for determinations of the sustainability of aquaculture products by retailers

fields of ecological engineering and industrial ecology. Indeed, tidal wetland, mangrove forest, coral, and sea grass restoration aquaculture – in addition to establishment and maintenance of oyster reefs – are important examples of aquaculture creating, enhancing, and maintaining productive marine ecosystems and habitats, and improving water quality.

### Aquaculture and Agriculture Science

Here, science questions as to whether aquaculture contributes to the depletion of world fisheries. Fed aquaculture depends on both wild and farmed fish stocks and on intact aquatic habitats and excellent water quality, plus a growing quantity of agricultural resources. There is much ongoing policy, research, and management concerns on the interactions of marine food fish fisheries (“biomass fisheries”) with aquaculture and human welfare. There is much less planning and research regarding the future impacts of fed aquaculture on agriculture.

Agricultural meals and oils as alternatives to marine sources are developing rapidly. Current projections forecast that fed aquaculture may use 50% or less of the world’s fish meal [32], which would mean a large expansion of use of agricultural and other terrestrial sources of feed proteins and oils. Terrestrial proteins and oils from soybeans, sunflowers, and lupins are available at volumes larger than the available global quantities of fish meal. Soybeans have high protein content of ~28%, peas have ~22%, and these have good amino acid profiles. Other abundant agricultural cereals have lower protein contents of ~12–15%. Processing can create protein concentrates with protein levels of >50% [33]. Vegetable oils have very low EPA (eicosapentaenoic acid) and DHA (docosahexaenoic acid) levels. However, substitution of plant oils upward of 50% of added dietary oil has not resulted in growth reductions or increased mortalities in fish such as salmon and trout.

If agricultural sources of meals and oils are the future of fed aquaculture, there will be a need for a new sustainability planning and science on the impacts of fed aquaculture as a driver of agriculture production, especially so for soybeans. Increased aquaculture consumption of the world’s grains and oils raises the concern over the spread of unsustainable agriculture

practices. Brazil has been targeted as one of the world’s major soybean suppliers. Costa et al. [34] have demonstrated that soybean farms are causing reduced rainfall in the Amazonian rainforest. About one seventh of the Brazilian rainforest has been cut for agriculture, about 15% of which is soybeans. Soybeans, which are light in color, reflect more solar radiation, heating the surface of the land less and reducing the amount of warm air convected from the ground. Fewer clouds form as a result, and less precipitation falls. In soybean areas, there was a 16% less rainfall compared to a 4% decrease in rainfall in land areas cleared for pasture.

### Aquaculture for the Poor

Approximately 1.3 billion people live on less than a dollar a day, and half of the world’s population lives on less than 2 dollars a day. FAO has stated that the world will need to produce 70% more food for an additional 2.3 billion people by 2050 [4]. Scarce natural resources will need to be used more efficiently, and there will be a need for proper socioeconomic frameworks to address imbalances and inequities to ensure that everyone in the world has access to the food they need. Food production will have to be carried out in a way that reduces poverty and takes account of natural resource limitations [4].

The world’s population will rise from 6.8 to 9.1 billion in 2050, with nearly all population growth occurring in the economically developing countries. Without additional global food strategies, an estimated 370 million people will be hungry in 2050. The magnitude of the problem is most acute in Africa. In 10 African countries of an estimated 316 million persons where aquatic proteins are an important dietary component, 216 million live on US\$2/day, 88 million are undernourished, and 16 million children under 5 are malnourished [35].

Small-scale coastal and inland freshwater fisheries provide more than 90% of the fish consumed in Africa. Over 2.5 million people are involved in fishing and 7.5 million in trading, marketing, and processing. The most important fisheries/aquaculture ecosystems are located on the coasts of west and southern Africa and the river basins of Senegal, Niger, Volta, Congo, Lake Chad, Nile, and Zambezi Rivers. But today, aquaculture provides less than 5% of Africa’s fish, with most concentrated in Egypt and Nigeria [35].

Aquaculture is a global enterprise with local roots. There are strong concerns that aquaculture is evolving away from its global responsibility to provide net benefits (additional foods) for a protein-hungry planet [36–38]. Greater than 75% of global fisheries are traded. In 2000, more than 60% of fish meal was traded. Only 7% of meat is traded, 17% of wheat, and 5% of rice. To tackle this huge challenge, the FAO ecosystems approach to aquaculture [39] has created a new code for responsible global aquaculture development, and has combined this into one common development framework for a global implementation strategy for aquaculture that can be used to measure the trajectory of social responsibility for global aquaculture.

If aquaculture is designed, implemented, and evaluated as aquaculture ecosystems, a new social contract would have a close relationship between aquaculture professionals who not only develop an alternative model of aquaculture development but also interact closely with capture fisheries and agriculture but help deliver to the world's poor its needs for nutrient dense, protein-rich seafoods. Components of a global strategy could be to:

1. **Allocate more food fish and oils for poverty alleviation and human needs worldwide, and allocate less marine resources for feed fish for fed aquaculture so as to: (a) increase the ecosystem resilience of the Humboldt ecosystem, and (b) relieve the increasing overdependence of aquaculture countries such as Thailand (shrimp) and Norway (salmon) on this southeastern Pacific Ocean marine ecosystem.**

Alder et al. [37] estimated that about 36% of the world's fisheries catch (30 million tons) are processed into fish meal and oil, mostly to feed farmed fish, chickens, and pigs. Daniel Pauly of the University of British Columbia has stated that "Globally, pigs and chickens alone consume six times the amount of seafood as US consumers and twice that of Japan." Jacquet et al. [28] reported that Peru exports about half of the world's fish meal from its catch of 5–10 MMT/year of anchovies while half of its population of 15 million live in poverty and 25% of its infants are malnourished.

A campaign launched in 2006 combining scientists, chefs, and politicians to demonstrate that anchovies are more valuable to the Peruvian people and its economy as direct foods has resulted in a 46% increase in demand fresh and 85% increase in canned anchovies. One ton of fillets has sold for five times the price of 1 t of meal and requires half the fish (3 t for 1 t fillets vs. 6 t for 1 t meal). Peru has decided to dedicate 30% of its annual food security budget (approx. US\$ 80 million) for programs to supply anchovies to its people. Higher prices for fish used as direct human foods for food security will limit processing of fish to meals for terrestrial animal and aquaculture feeds, thereby decreasing the supply of fish meals and oils for global aquaculture trade and development, but meeting the Millennium Development Goals of eliminating everywhere extreme hunger and starvation.

2. **Accelerate research into the elucidating functional feed ingredients in fish diets that are showing the potential to eliminate the needs for fish meal and oils in aquaculture.**

Skretting Aquaculture Research Centre [40] reported a research on "functional ingredients" that are contained in fish meals and oils, which contribute to efficient feed conversions and high growth rates, fish health, and welfare. Initial research focused on beta-glucans that not only stimulate the immune system of fish and protect against the effects of bacterial furunculosis but also allow reductions in fish meal contents in diets to 25%. Additional research with phospholipids in meals, triglycerides in fish oil, and antioxidants in 2008 have resulted in excellent fish performances from feeds with almost no marine fish meal and oil. Current research is exploring the extraction of functional ingredients from other non-marine by-products.

3. **Develop alternative ecological aquaculture models that accelerate the movement toward use of agricultural, algal, bacterial, yeasts meals, and oils.**

Aquaculture uses most of the world's fish meal (68%) and fish oil (88%); however, Tacon and Metian [32] predict that fish meal and oil use in aquaculture will decrease to become high priced,

specialty feed ingredients. Currently, about 40% of aquaculture depends on formulated feeds: 100% of salmon, 83% of shrimp, and 38% of carp. As stated previously, research on the use of agricultural meals and oils to replace use of ocean resources especially on the functional components of fish meals/oils needed for fish nutrition is a major subject of aquaculture research and development [41, 42]. Turchini et al. [43] reported that for all of the major aquaculture fish species that 60–75% of dietary fish oil can be substituted with alternative lipid sources without significantly affecting growth performance, feed efficiency, and feed intake. Naing et al. [44] found that palm oil could replace fish oil in rainbow trout diets, and reduce the dioxin contents in fish.

**4. Develop new governance systems that integrate aquaculture, agriculture, and fisheries using ecosystem-based management approaches, which combine production, distribution, and consumption networks that do not institutionalize poverty and hunger, but provide new alternative tools and education in multisectoral ecosystem approaches.**

The massive environmental change being brought about by the accelerated growth of the world's population has caused profound change to the world's ecosystems. Crutzen and Stoermer [45] have called this new era the "Anthropocene." In this era, massive quantities of additional foodstuffs will be needed to sustain humanity; nutrient-dense, high-quality aquatic proteins will be especially important. The tools and training for creating the next generation of transdisciplinary, sustainability scientists will need to be further developed and well utilized; otherwise, it will result in serious consequences for the Earth's living systems.

## Bibliography

### Primary Literature

- Ostrom E (1986) An agenda for the study of institutions. *Public Choice* 48(1):3–25
- Costa-Pierce B (2002) *Ecological aquaculture*. Blackwell, Oxford
- Costa-Pierce B (2008) An ecosystem approach to marine aquaculture: a global review. In: Soto D (ed) *Building an ecosystem approach to aquaculture*, pp 81–116. FAO Fisheries and Aquaculture Proceedings 14, FAO, Rome, 221p
- FAO (Food and Agriculture Organization) (2009) The state of world fisheries and aquaculture 2008. Fisheries Department, Food and Agriculture Organization of the United Nations, Rome
- FAO (1995) Code of conduct for responsible fisheries. FAO, Rome, 41 pp. Available at <ftp://ftp.fao.org/docrep/fao/003/W4493e/W4493e00.pdf>
- Moore B III, Underdal A, Lemke P, Loreau M (2002) The Amsterdam declaration on global change. In: Steffen W, Jäger J, Carson DJ, Bradshaw C (eds) *Challenges of a changing earth*. Springer, Heidelberg
- Gliessman S (1998) *Agroecology: ecological processes in sustainable agriculture*. Ann Arbor Press, Ann Arbor
- Altieri M (2002) Agroecology: the science of natural resource management for poor farmers in marginal environments. *Agr Ecosyst Environ* 93:1–24
- Kates R, Clark W, Corell R (2001) Sustainability science. *Science* 292:641–642
- Wurts W (2000) Sustainable aquaculture in the twenty-first century. *Rev Fish Sci* 8:141–150
- Olsen S, Sutinen J, Juda L, Hennessey T, Grigalunas T (2006) A handbook on governance and socioeconomics of large marine ecosystems. Coastal Resources Center, University of Rhode Island, Narragansett
- Olsen S (2003) Frameworks and indicators for assessing progress in integrated coastal management initiatives. *Ocean Coast Manage* 46(3–4):347–361
- Olsen S, Page G, Ochoa E (2009) The analysis of governance responses to ecosystem change: a handbook for assembling a baseline. LOICZ Reports & Studies No. 34. GKSS Research Center, Geesthacht, Germany
- GESAMP (IMO/FAO/UNESCO-IOC/UNIDO/WMO/IAEA/UN/UNEP Joint Group of Experts on Scientific Aspects of Marine Environmental Protection) (2008) Assessment and communication of environmental risks in coastal aquaculture. FAO, Reports and Studies GESAMP No. 76, Rome, 198 pp
- Tucker C, Hargreaves J (2008) *Environmental best management practices for aquaculture*. Wiley-Blackwell, Ames, IA
- National Research Council (2010) *Ecosystem concepts for sustainable bivalve mariculture*. National Academies Press, Research Council, Washington, DC
- Boyd C, Tucker C, McNevin A, Bostick K, Clay J (2007) Indicators of resource use efficiency and environmental performance in fish and crustacean aquaculture. *Rev Fish Sci* 15:327–360
- Daly HE (1996) *Beyond growth: the economics of sustainable development*. Beacon, Boston
- Pullin R (1993) Discussion and recommendations on aquaculture and the environment in developing countries. In: Pullin R, Rosenthal H, Maclean J (eds) *Environment and aquaculture in developing countries*. ICLARM Conference Proceedings 31, pp 312–338

20. Hephher B (1985) Aquaculture intensification under land and water limitations. *GeoJournal* 10:253–259
21. Mires D (2009) Development of inland aquaculture in arid climates: water utilization strategies applied in Israel. *Fisheries Manag Ecol* 7:189–195
22. Neori A, Shpigel M, Ben-Ezra D (2000) A sustainable integrated system for culture of fish, seaweed and abalone. *Aquaculture* 186:279–291
23. Pimentel D, Pimentel M (2003) Sustainability of meat-based and plant-based diets and the environment. *Am J Clin Nut* 78(3):660–663
24. Fletcher W, Chesson J, Fisher M, Sainsbury K, Hundloe T (2004) National ESD reporting framework: the 'how to' guide for aquaculture. FRDC Project 2000/145.1, Canberra, 88 pp
25. FAO Fisheries and Aquaculture Department (2006) State of the world's fisheries and aquaculture. FAO Italy, Rome, 162 p. Available at <http://www.fao.org/docrep/009/A0699e/A0699E00.htm#Contents>
26. Cadenasso M, Pickett S, Grove M (2006) Integrative approaches to investigating human-natural systems: the Baltimore ecosystem study. *Nat Sci Soc* 14:4–14
27. Roheim C (2009) An evaluation of sustainable seafood guides: implications for environmental groups and the seafood industry. *Mar Resour Econ* 24:301–310
28. Jacquet J, Hocevar J, Lai S, Majluf P, Pelletier N, Pitcher T, Sala S, Sumaila R, Pauly D (2009) Conserving wild fish in a sea of market-based efforts. *Oryx, The Int J Cons.* doi:10.1017/S0030605309990470
29. Roheim C (2001) Product certification and ecolabelling for fisheries sustainability FAO Fisheries Technical Paper 422. FAO, Rome, Italy
30. Wessells C, Johnston R, Donath H (1999) Assessing consumer preferences for ecolabelled seafood: the influence of species, certifier, and household attributes. *Am J Agr Econ* 5: 1084–1089
31. Costa-Pierce B, Bridger C (2002) The role of marine aquaculture facilities as habitats and ecosystems. In: Stickney R, McVey J (eds) *Responsible marine aquaculture*. CABI, Wallingford, pp 105–144
32. Tacon A, Metian M (2008) Global overview on the use of fish meal and fish oil in industrially compounded aquafeeds: trends and future prospects. *Aquaculture* 285:146–158
33. Bell J, Waagbo R (2002) Safe and nutritious aquaculture produce: benefits and risks of alternative sustainable aquafeeds. In: Holmer M, Black K, Duarte C, Marba N, Karakasis I (eds) *Aquaculture in the ecosystem*. Springer, New York, pp 185–226
34. Costa M, Yanagi S, Souza P, Ribeiro A, Rocha E (2007) Climate change in Amazonia caused by soybean cropland expansion, as compared to caused by pastureland expansion. *Geophys Res Lett* 34(L07706). doi:10.1029/2007GL029271
35. Allison E, Beveridge M, van Brakel (2009) Climate change, small-scale fisheries and smallholder aquaculture. In: Culberg M (ed) *Fish, trade and development*. Royal Swedish Academy of Agriculture and Forestry, Stockholm, pp 73–87
36. Goldburg R, Naylor R (2005) Future seascapes, fishing, and fish farming. *Front Ecol Environ* 3(1):21–28
37. Alder J, Campbell B, Karpouzi V, Kaschner K, Pauly D (2008) Forage fish: from ecosystems to markets. *Ann Rev Environ Res* 33:71–714
38. Naylor R, Hardy R, Bureau D, Chiu A, Elliott M, Farrell A, Forster I, Gatlin D, Goldburg R, Hua K, Nichols P (2009) Feeding aquaculture in an era of finite resources. *Proc Nat Acad Sci* 106(36): 15103–15110
39. Soto D, 21 co-authors (2008) Applying an ecosystem-based approach to aquaculture: principles, scales and some management measures. In: Soto D, Aguilar-Manjarrez J, Hishamunda N (eds) *Building an ecosystem approach to aquaculture*. Rome: FAO Fisheries and Aquaculture Proceedings 14. FAO, Rome, 221 pp
40. Skretting Aquaculture Research Centre (2009) Revealing the essential functions of vital feed ingredients. Accessed at <http://www.skretting.com/Internet/SkrettingGlobal/webInternet.nsf/wPrid/D30B62DD93F65CFEC1257663004DA962!OpenDocument>
41. Watanabe T (2002) Strategies for further development of aquatic feeds. *Fish Sci* 68:242–252
42. Opstvedt J, Aksnes A, Hope B, Pike I (2003) Efficiency of feed utilization in Atlantic salmon (*Salmo salar* L.) fed diets with increasing substitution of fish meal with vegetable proteins. *Aquaculture* 221:365–379
43. Turchini G, Bente E, Torstensen E, Ng W (2009) Fish oil replacement in finfish nutrition. *Rev Aquacult* 1(1):10–57
44. Naing A, Satoh S, Tsuchida N (2007) Effect of replacements of fishmeal and fish oil on growth and dioxin contents of rainbow trout. *Fish Sci* 73:750–759
45. Crutzen P, Stoermer E (2000) The "Anthropocene". *Glob Change Newslett* 41:17–18
46. Fletcher A (2003) Mapping stakeholder perceptions for a third sector organization. *J Intell Cap* 4(4):505–527
47. Savage G, Nix T, Whitehead B (1991) Strategies for assessing and managing organizational stakeholders. *Acad Manage Exec* 5(2):61–75
48. Hemmati M, Dodds F, Enayti J, McHarry J (2002) Multistakeholder processes on governance and sustainability. Earthscan, London
49. Dalton T (2005) Beyond biogeography: a framework for involving the public in planning of U.S. Marine Protected Areas. *Conserv Biol* 19:1392–1401
50. Dalton T (2006) Exploring participants' views of participatory coastal and marine resource management processes. *Coast Manage* 34:351–367
51. ISSD (International Institute for Sustainable Development) (2004) ISO social responsibility standardization. An outline of the issues. International Institute for Sustainable



Development, Winnipeg, Manitoba. [http://www.iisd.org/pdf/2004/standards\\_iso\\_srs.pdf](http://www.iisd.org/pdf/2004/standards_iso_srs.pdf)

52. <http://www.iclei.org>
53. American Center for Life Cycle Assessment. <http://www.lcacenter.org>
54. Bartley D, Brugère C, Soto D, Gerber P, Harvey B (2007) Comparative assessment of the environmental costs of aquaculture and other food production sectors: methods for meaningful comparisons. FAO/WFT Expert Workshop, 24–28 April 2006, Vancouver, Canada. FAO Fisheries Proceedings. No. 10. FAO, Rome, 241p
55. Ayer N, Tyedmers P (2009) Assessing alternative aquaculture technologies: life cycle assessment of salmonid culture systems in Canada. *J Cleaner Prod* 17:362–373
56. International Organization for Standardization (ISO). [http://www.iso.org/iso/iso\\_14000\\_essentials](http://www.iso.org/iso/iso_14000_essentials)
57. Environmental Impact Assessment Review. [http://www.elsevier.com/wps/find/journaldescription.cws\\_home/505718/description#description](http://www.elsevier.com/wps/find/journaldescription.cws_home/505718/description#description)
58. IAIA (International Association for Impact Assessment) (1999) Principles of environmental impact assessment best practices. IAIA, Fargo, ND, USA. [http://www.iaia.org/publicdocuments/special-publications/Principles%20of%20IA\\_web.pdf](http://www.iaia.org/publicdocuments/special-publications/Principles%20of%20IA_web.pdf)
59. US Department of Transportation. [http://www.fhwa.dot.gov/planning/toolbox/costbenefit\\_forecasting.htm](http://www.fhwa.dot.gov/planning/toolbox/costbenefit_forecasting.htm)
60. Savitz A (2006) The triple bottom line: how today's best-run companies are achieving economic, social and environmental success – and how you can too. Jossey-Bass/Wiley, New York
61. McCandless M, Venema H, Barg S, Osborne B (2008) Full cost accounting for agriculture – final report. Valuing public benefits accruing from agricultural beneficial management practices: an impact pathway analysis for Tobacco Creek, Manitoba. International Institute for Sustainable Development, Winnipeg, Manitoba, Canada. [http://www.iisd.org/pdf/2008/measure\\_fca\\_2008.pdf](http://www.iisd.org/pdf/2008/measure_fca_2008.pdf)
62. Gibson R, Hassan S, Holtz S, Tansey J, Whitelaw G (2005) Sustainability assessment: criteria and processes. Earthscan, London

## Books and Reviews

- Pullin R (1993) Environment and aquaculture in developing countries. International Center for Living Aquatic Resources Management, Manila
- Smil V (2000) Feeding the world: a challenge for the twenty-first century. MIT Press, Cambridge MA
- Tacon A, Hasan MR, Subasinghe RP (2006) Use of fishery resources as feed inputs for aquaculture development: trends and policy implications. FAO Fisheries Circular No. 1018. Food and Agriculture Organization of the United Nations, Rome

## Aquapod Systems for Sustainable Ocean Aquaculture

STEPHEN H. PAGE

Ocean Farm Technologies, Inc., Morrill, ME, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Aquapod Technology and Engineering Description

Aquapod Installations and Experience to Date

Future Directions

Bibliography

### Glossary

**Artisanal fisheries** Traditional fisheries involving fishing households (as opposed to commercial companies), using relatively small amount of capital and energy, relatively small fishing vessels or canoes, often beach-based, making short fishing trips, close to shore, mainly for local consumption.

**Biofouling** The accumulation of living organisms on some surface by bacteria, fungi, protozoa, algae, and invertebrates.

**Geodesic** The shortest line between two points on a specific surface.

**Open ocean aquaculture** The culture of marine organisms in exposed ocean locations, not sheltered by islands or embayments.

### Definition of the Subject

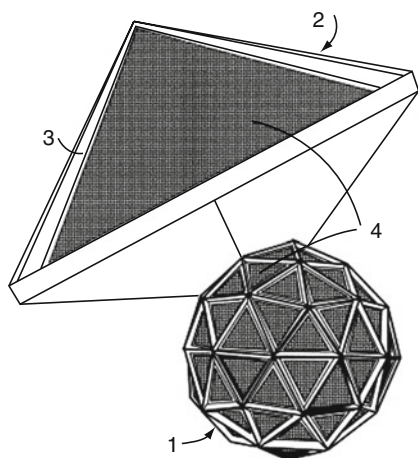
Open ocean aquaculture is not well defined by the industry, but in general refers to culture of marine fish, invertebrates, or algae in exposed ocean locations. Open ocean aquaculture is contrasted to near-shore marine aquaculture in that it occurs in areas removed spatially from land, typically by 1 km or more; deep water, generally deeper than 20 m; and exposure to wind, waves, and ocean currents without shelter from the mainland or islands. The subject of this entry is

principally the design and engineering of containment systems suited for open ocean aquaculture.

## Introduction

Near-shore finfish aquaculture worldwide is challenged and constrained by resource user competition, environmental carrying capacity of near-shore environments, and, in most cold-water regions of the world essentially a monoculture of Atlantic salmon. Expansion of marine aquaculture in the next 20 years will happen substantially in offshore, exposed open ocean areas, with a diversification of species.

The patented Aquapod™ is a unique containment system for marine aquaculture, suited for rough open ocean conditions and a diversity of species. The Aquapod is constructed of individual triangular net panels fastened together in a spheroid shape (Fig. 1). The majority of the panels are simply structural members and netting. Some individual panels or groups of



**Aquapod Systems for Sustainable Ocean Aquaculture. Figure 1**

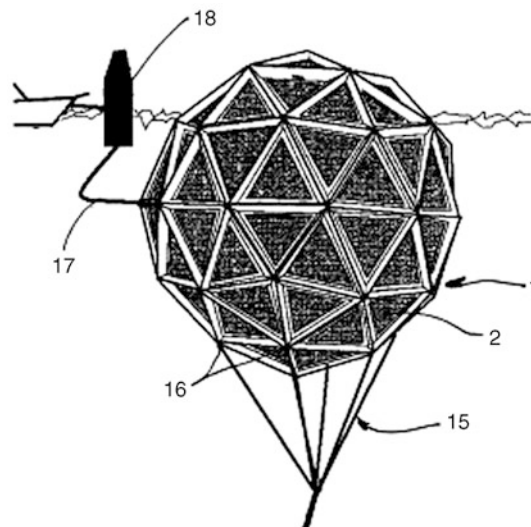
The scale of contemporary net pens to date has become problematic especially in terms of net-handling activities, which are labor intensive, require significantly sized hydraulic equipment, and cannot be performed in rough ocean conditions. The modular nature of the Aquapod solves this problem by utilizing individual net panels which are interchangeable and scaled to be handled without difficulty in adverse conditions. These net panels are fastened together to form a secure enclosure, tailored to each customer's specific requirements

panels have other functions, such as access, feeding, fish transfer and grading, harvest, mooring, and mortality recovery. Other individual panels may have pneumatically controlled flotation devices which allow an almost infinite orientation of the Aquapod in the water. The Aquapod functions as a total containment system for finfish while submerged or partially surfaced.

Elements of the design have benefits for reduction in labor for routine husbandry tasks, reduction in maintenance costs, and reduction of stress on aquatic animals during handling operations such as transfer, treatment, and harvest. The "exoskeleton" design of the Aquapod containment system also allows for internal structures to provide for the cultivation of flatfish species such as halibut and flounder.

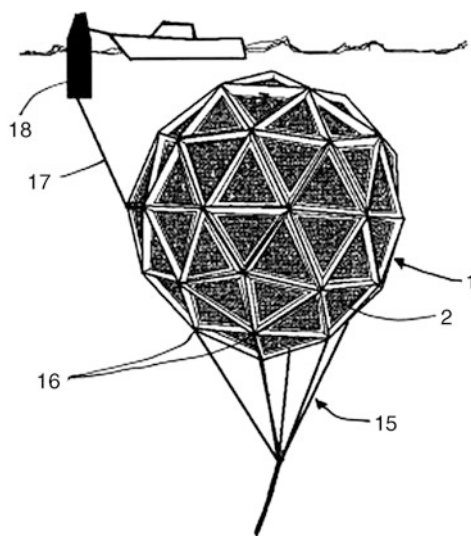
## Features of the Aquapod Containment System

**Submersibility:** Submersion is the preferred, if not the only way to operate fish containment systems in the open ocean. Submergence is necessary to operate a fish pen below the destructive energy of surface waves. Submergence also allows fish to be kept at favorable temperatures below warm water thermoclines. The Aquapod can be operated partially surfaced (Fig. 2)



**Aquapod Systems for Sustainable Ocean Aquaculture. Figure 2**

The structure of the Aquapod net pen allows for many configurations of mooring attachment. Figure 2 shows an Aquapod net pen partially surfaced



### Aquapod Systems for Sustainable Ocean Aquaculture.

**Figure 3**

An Aquapod net pen submerged

or fully submerged (Fig. 3). Although there are other pen designs which are designed to operate submerged, the spherical Aquapod design is well suited for species such as Atlantic salmon, which require periodic access to substantial surface area in order to gulp air.

**Fixed volume:** The Aquapod maintains its shape and volume in strong currents or undertow, reducing stress and physical damage to the fish contained within. Aquapod net pens can be constructed to any size to fit individual customer's needs, with a practical range of 8 m diameter (212 m<sup>3</sup>) to 24 m diameter (7,000 m<sup>3</sup>).

**Containment:** The structural exoskeleton of the Aquapod provides a high ratio of structural support to net area. Each triangular panel is comprised of an essentially rigid frame supporting tensioned netting. Frames can be constructed of marine grade aluminum, fiber-reinforced polymer, or even a combination of materials. The modular nature of the Aquapod containment system provides for easy inspection and inventory control as required by today's mandated containment management systems. Individual net panels are coded so they can be brought into a regular schedule of inspection, maintenance, repair, and replacement. While surfaced in suitable conditions, a single net panel or a group of net panels can be removed and replaced without compromising the

integrity of the containment. No other system can offer this convenience.

**Predator control and escapement:** Tensioned netting has proven to be relatively predator proof in other engineered systems. The Aquapod goes one step further. Due to the modular structure and the relatively small triangular units, different types of wire mesh can be used in place of synthetic netting. This enhances predator control in areas where triggerfish, sharks, sea lions, and crocodiles have made fish farming impractical. Wire mesh also prevents escapement by chewing fish such as cod and sea bream, which readily chew through synthetic fiber netting.

**Husbandry and fish health:** Fish can be moved from one Aquapod containment to another without seining or pumping. Patent pending displacement technology is capable of transferring fish from one Aquapod containment pen to another without brailing or pumping, or from an Aquapod containment to a harvesting pump without seining or other stressful means of crowding the fish. While being transferred, fish can be graded and counted with automated technology. In competitive designs, these fish-handling operations are either impossible or very labor intensive and stressful on the fish.

**Feeding:** One or more panels in each Aquapod containment is modified to receive and distribute hydraulically supplied feed from a centralized feed barge or service boat. The semirigid and modular exoskeleton of the Aquapod allows easy attachment of any number of feeding ports. Multiple feed outlets provide better distribution of feed to the fish.

**Safety:** Worker safety is a paramount concern when operating fish farms in open ocean conditions. Since conception, the Aquapod containment system has been designed to maximize automation of routine husbandry tasks and reduce the amount of time divers are needed in the water. Although any containment system will need some diving, the ability of the Aquapod to rotate within its mooring grid, bringing any segment of the pen to the surface or near to the surface, reduces the amount of diving, and when diving is needed, the depth of the dives will be greatly reduced. One of the most dangerous jobs on a fish farm is net changing, and the Aquapod eliminates this chore by providing a means to clean nets at the surface.

**Waste management:** The spherical design of the structure causes mortalities to collect at the bottom of the pen where they can be brought to the surface with a conventional airlift pump, eliminating the routine and hazardous task of mortality collection by divers (Fig. 4).

**Mooring.** There are many options for mooring an Aquapod net pen. Each hub of the frame is a potential mooring point, which allows enormous flexibility in mooring placement and the ability to distribute loads over a large area for safety. Any specific net panel can be strengthened to facilitate predicted maximum mooring stresses, whether the Aquapod is attached to a single-point mooring or whether it is installed in a conventional submerged grid system (Fig. 5). Another mooring option available in suitable site is a two-point mooring (Fig. 6) that facilitates Aquapod rotation.

**Cost:** On a capital cost per cubic meter of containment basis, the Aquapod containment system is significantly more than conventional surface pens but significantly less than the cost of currently available submersible net pens. Furthermore, when submerged, the volume used to calculate containment is accurately figured, unlike surface pens which calculate the volume

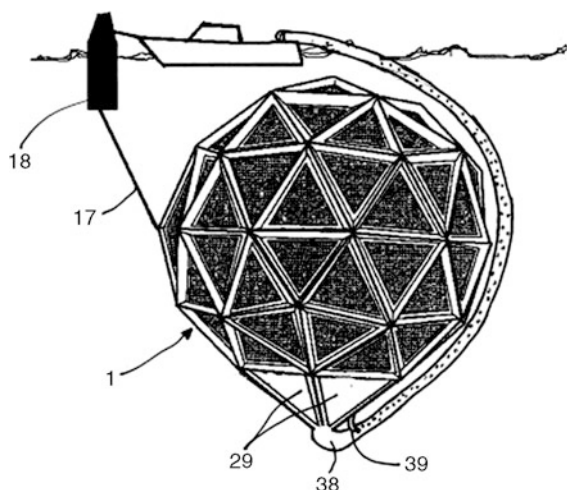
in the top part of the water column, although that space is not used by the fish. Operational costs are less than existing systems due to the efficiency of the proprietary fish transfer technology for routine husbandry tasks. Maintenance costs are reduced by the modular nature of the net panels and the ability to orient the Aquapod so that any part of the net pen is at or near the water surface, making it easy to inspect, remove, and replace individual net panels. When optional vinyl-coated wire mesh is used for netting, net replacement is reduced and net washing is greatly facilitated.

The true cost of any system is not only the initial capital cost, but the life cycle cost and the cost of operations, including risk. The Aquapod has been designed with these factors as principal drivers.

## Aquapod Technology and Engineering Description

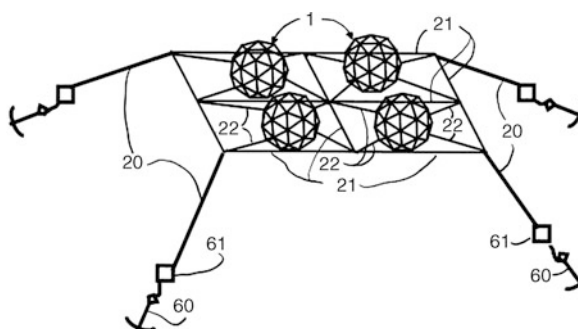
### Current State of Design and Materials

The engineering work to date on the Aquapod design has centered primarily on early 10 m (31 ft) diameter units, the larger 19 m (64 ft) diameter A3600 size, and more recent 8 m (27 ft) A212 Micropods. All these models have used recycled HDPE materials as the primary element for the struts, and plastic-coated welded wire mesh for the net fabric. Selective reinforcement around the mooring attachment points has addressed strength issues. So far these materials seem to work well and have been a basis for all Aquapod installations to date. The innovations by Ocean Farm Technologies



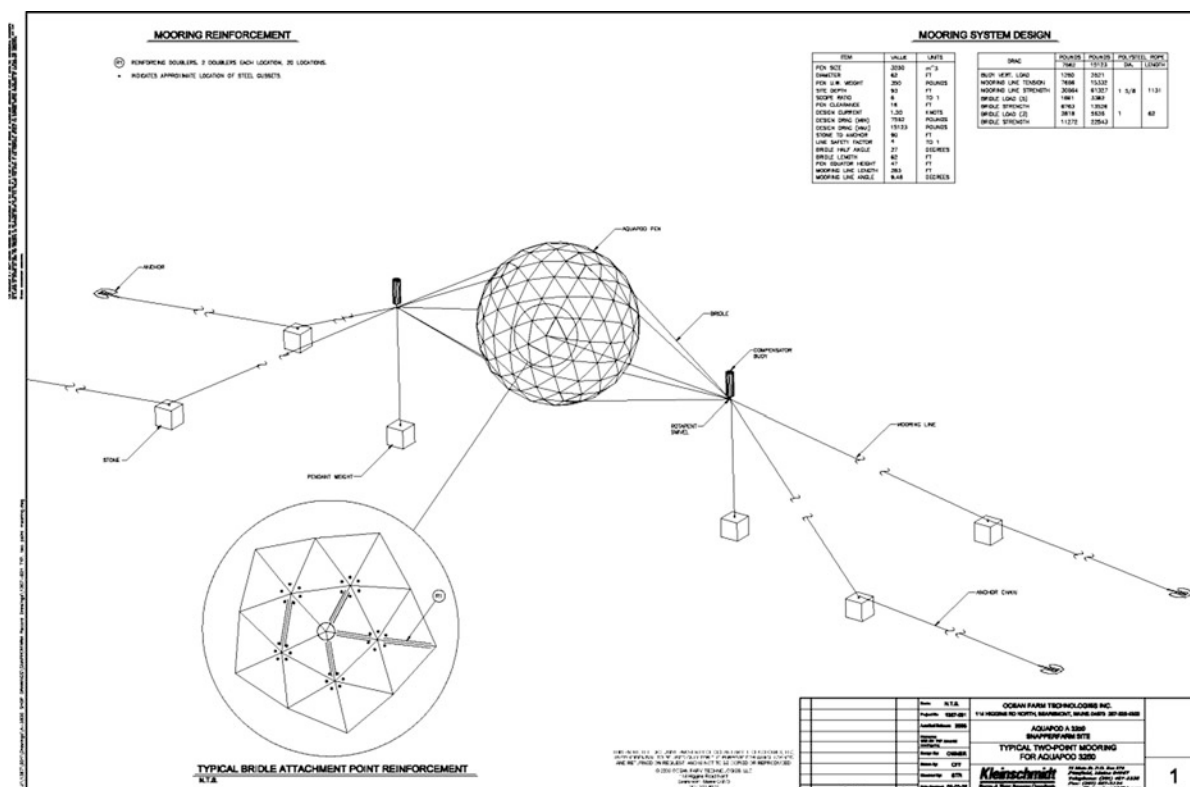
**Aquapod Systems for Sustainable Ocean Aquaculture.**  
**Figure 4**

A simple air lift suction pump can remove mortalities and waste from the bottom of the Aquapod net pen for removal and processing



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
**Figure 5**

Aquapod net pens can be moored in many different configurations of grids



Aquapod Systems for Sustainable Ocean Aquaculture. Figure 6

Single Aquapod net pens can be moored individually in a two-point mooring. With swivels at the bridle attachment points, this allows for rotation of the net pen

(OFT) with testing and analysis have been steadily advancing the design, learning the relative cost and difficulty of various ideas, with the goal of refining the concept and improving the economy of the design, manufacture assembly, and installation procedures.

Mooring system components have been selected to create a two-point design allowing movement in the vertical direction and rotation, and single-point moorings. Buoyancy calculations have been done to confirm the near-neutral buoyancy of the system.

The main engineering goal is to design an economical pen that is strong enough, but not excessively heavy or expensive, by defining the real safety factor of the current design, and seeing where further structural design improvements might be made. With a large structure with so many repetitive elements, the changes in any one element are repeated hundreds of times, and thus has a real effect on cost and fabrication. Further work is also being done by OFT with more long-range

improvements and concepts for different types of designs using different materials or assembly ideas.

Further data from the real size pens has been used to confirm drag forces. Loading of the pen has to date been based on a tow test performed by a University of New Hampshire boat on the 31-ft geodesic pen and has been scaled up for the 62-ft pen [1]. OFT plans to measure drag loads on existing pens as a function of current velocity and thus more accurately establish this relationship. The current drag forces used to date are on the conservative side, allowing for the effect of biofouling that increases drag.

The structural effect of wire mesh on the Aquapod is obvious but difficult to measure or model. The mesh contribution has not been credited in the analysis to date, except to provide buckling restraint for the compression member weak axes. As far as basic strength, the mesh can help reduce tension stresses, but will do nothing to reduce compressive stresses in the struts,



and since these are usually of about the same magnitude over the sphere, the overall effect may not be of much importance.

### Results of University of Maine Engineering Study

The use of recycled HDPE [2] plastic extrusions in this type of structural engineering application is rare and, for the designer, the available data on materials and fasteners in the material is limited. For design of the Aquapod struts, it is necessary to check the capacity of the tension and compression members. Needed for compression members are values to use for modulus of elasticity, maximum allowable stresses for compression, based on buckling, yielding, and crushing, and creep rupture of the real life struts. For design of the tension members, the modulus of elasticity, maximum and allowable tensile stresses, and fastener strengths are needed. In addition, bolts and other fasteners are used in the Aquapod and data is needed regarding end and edge distances, strengths, and yield points. Testing of the full size members and hub assemblies has been desirable to confirm the actual strength of real life members, and thus provide an indication of the real safety factor provided in the pen.

The Hybrid Structures Testing Laboratory in the Mechanical Engineering Department, at the University of Maine, was engaged by OFT and performed initial testing (1) to determine engineering data for the plastic material that is used in the Aquapod net pen system and (2) to test sub-component details of the Aquapod system. The goal has been to have design guidance to use as the limits for the members, as calculated for the pen under the various load scenarios. Testing was performed under the direction of Vincent Caccese, Ph.D., P.E., Professor of Mechanical Engineering (Caccese).

The testing consisted of a total of 62 structural tests of components as follows:

1. Material Bolt Bearing, Tensile, and Shear-out Capacity (40 tests performed)
2. Column Compression Tests (20 tests performed)
3. Mesh Assembly Buckling (2 tests performed)

The bolt tests were done first and arrived early enough to size the bolts in the Aquapod 3250. These tests were primarily to determine the bolt capacities for the

tension members in the Aquapod, and to determine how many, and what size bolts would provide the required tension transfer from component to component, principally in the steel-reinforced connections around the mooring attachment points. These tests also showed that the tensile capacity of the struts is a controlled by the holes, and tension members will fail first at the bolt holes.

The compression tests were primarily aimed at determining maximum compression loads that could be assigned to struts, which should be a safe percentage of the ultimate compression demonstrated by the full size tests, and should also be a safe amount below the maximum stress that induces the beginning of buckling.

While the sample sizes for the compression tests were small, the results indicate that the actual modulus of elasticity for design use is significantly higher than the published data provided by the plastic vendor for design of compression posts, but well under the value listed as an engineering property in their material property tables. The actual reinforcement of members with doublers could have been reduced somewhat and the current design is conservatively reinforced.

### Assessment of Structural Stability of Aquapod Net Pens

The geodesic sphere that has been used for this pen is based on an icosahedron. The icosahedron is a geometric shape consisting of ten triangular faces that are further subdivided into triangular panels, in the process of creating the geodesic sphere. Each of the ten faces is subdivided into 16 triangular panels of 10 ft on a side. Each strut is a pair of extruded fiberglass-reinforced HDPE members. All panels are faced with 1" × 1" × 16" gage vinyl-coated galvanized steel wire mesh.

The geodesic sphere has been structurally analyzed using finite element computer modeling, for load cases due to various loading conditions that might be encountered in the life of a pen. Under each load case, an appropriate safety factor related to the likelihood of the case was used. The load cases are:

- Drag force from current at the site or when being towed. Using a total drag force that was extrapolated from the tow testing results, the total force is

distributed around the pen based on locating maximum forces on the leading and trailing faces, as oriented to the current.

- Dead load of a pen hanging from a crane on its five part bridle.
- Drag forces with the pen in a current but with one and two broken bridle lines.

There have been numerous numerical simulations made of mooring line arrangements to refine the mooring system arrangement, and this has been an iterative process of examining the forces in the pen from a given loading and mooring arrangement. Through examination and rejection of a number of possible mooring schemes, OFT was able to come up with the final bridle arrangement that minimizes and localizes the maximum loading in the struts. This arrangement features five tangent bridle lines attached at the intersections of the base icosahedron faces, which, in the pen, are seen as a five panel intersection. The five panels form a pentagon. The analysis has found that locating the mooring at this position and angle will reduce the maximum strut forces and provide a superior distribution of loads within the structure.

The hub reinforcement consists of bent steel gusset replacements for the plastic gussets in each corner of the panels framing into the pentagons, in effect creating a steel ring at each of these hubs. The central hub of each pentagon receives a steel fabrication that reinforces the hub and provides an attachment point for the mooring bridle. The design for the central hub was developed for the conservative a 2-knot current loading, and a lighter design was developed for the Beta pen where the current is less.

When the member strut forces from each load case were sorted by magnitude, they were compared to critical allowable loading, assigned to the compression or tension members. When mapped to the actual positions, in all cases, the maximum strut loads, both tension and compression, are relatively concentrated in the pentagons around the mooring forces, and a few panels around them. The geodesic sphere distributes loading very quickly away from these areas and the struts forces become very low. Thus, OFT developed the recommended reinforcement of the pentagons related to the mooring attachments. This consisted of

selective triangle receiving steel gussets and some additional bolting.

In addition, OFT looked at the sphere under the same drag forces but with various broken tension or compression struts, and did not find much difference as compared to a broken bridle line. Critical strut failure, next to a mooring attachment, at most will relieve the mooring force of that line and redistribute load to the others. Failure of struts elsewhere in the sphere just redistributes around the broken strut area to adjacent struts, which have plenty of reserve. Thus, the structure appears to be very robust and redundant.

OFT has engineered on the conservative side with reinforcement requirements. The total test compressive load of 4,800 lbs on the unmeshed strut is more than the load previously calculated using the material properties provided by the vendor, and the load with the mesh included is even greater. The design has been using 3,800 lbs for the buckling strength of the short sides of the mooring panels, which is quite a bit less than either of the test results. The test data shows that the compressive buckling modulus of elasticity from the published value of the vendor is too low, since the test buckling strength is several times higher than the predicted strength. In the A3250, all members with compression loads greater than their buckling load calculated with  $E = 75,000$  were reinforced. Since the tests show greater  $E$ , around 150,000, this reinforcement is quite conservative.

### Feasibility of Scale-Up

OFT performed a preliminary analysis of a 28-m diameter pen using a scaled-up version of the A3250. This is an icosahedron but uses many additional panels of the same basic size of about 10 ft to a side, and encloses  $11,000 \text{ m}^3$ . The dead load case is the most severe. This is the load seen if the sphere were hung from a crane. The drag load with one broken mooring line appears to cause a similar magnitude axial loading in the worst struts. Since the dead load is 90,500 lbs., it will exceed the drag load determined using the drag calculation. The drag at 1 m/s current (2 knots) would be around 60,000 lbs.

The most-loaded members are all associated with the mooring line attachment points as before. The pentagon at each mooring line is highly loaded, more so than the previous A3250 aquapod, and will be

specifically reinforced as before. These loadings are relatively discrete areas and an inexpensive reinforcement will be devised for these pentagons, or they will be specially fabricated from a different material or configuration.

There are also high loads in the next members radiating from the pentagon on the downstream side. After the pentagon, most of the highest stressed members are in the hexagon of six panels just downstream of the pentagon. This hexagon will be treated like the pentagon and reinforced to handle the loads. The rest of the pen seems to have relatively lower loads. The extruded HDPE material with wire mesh is very capable of these loadings.

### Aquapod Installations and Experience to Date

Ocean Farm Technologies has installed 26 Aquapod net pens prior to 2011. Initial prototypes were deployed in Maine and New Hampshire (Figs. 7–9). The first commercial-sized installation of a Beta model was in Culebra Puerto Rico at Snapperfarm Inc (Fig. 10). A commercial crop of cobia (*Rachycentron canadum*) was grown in this Aquapod. OFT proceeded to develop small net pens with installations in Panama (Figs. 11, 12 and 13) and Mexico 9 (Fig. 14). The first commercial sale, in 2008, was to a company in South Korea which was in transition from wild-catch fishery to farming (Figs. 15 and 16). Most recent sales have been to

Mexico, where several sizes of Aquapods are deployed to grow shrimp (*p. vannamei*) (Figs. 17, 18, and 19).

### Future Directions

#### Low-Volume High-Density Net Pen Culture for Open Ocean Aquaculture

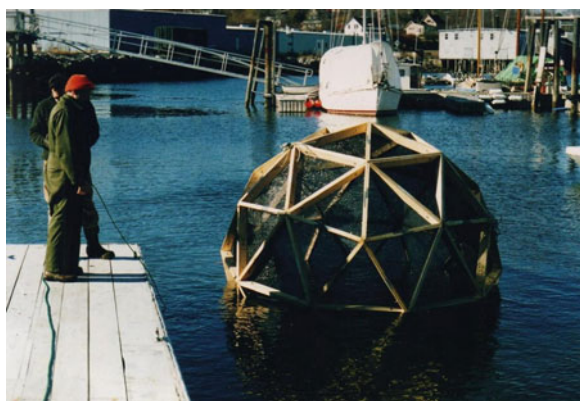
The emergence of an open ocean aquaculture industry provides an opportunity to reexamine traditional practices, such as the preference for ever-larger net pens for fish containment. The industry's focus on maximizing



Aquapod Systems for Sustainable Ocean Aquaculture.

Figure 8

Bucks Harbor, Maine 2005 prototype B A400



Aquapod Systems for Sustainable Ocean Aquaculture.

Figure 7

Belfast, Maine, USA 2004 prototype A



Aquapod Systems for Sustainable Ocean Aquaculture.

Figure 9

Portsmouth, New Hampshire, USA 2005 prototype C A400



net pen size is driven by the convention that large net pens provide more growing volume for a given investment. Because net pen costs can vary with the square of linear dimension while the volume varies by the cube, this assumption has merit. However, the assumption that larger net pens are more cost effective requires that stocking density remain unchanged throughout the size comparison. Stocking density in terms of kilograms per cubic meter is a useful fish-husbandry parameter in pond- and tank-based aquaculture

where extensive research has been done and where water exchanges are predetermined. There is also a growing body of data regarding stocking density in salmon net pens; however, these studies have been done in near-shore, low-energy environments. Little or no research has been done in open ocean net pen culture comparing density optimization over a range of net pen volumes, especially small net pens.

The principle advantage of small pens is that on-growing fish are closer to their source of clean



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
**Figure 10**  
Culebra, Puerto Rico 2006 Beta 1 A3250



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
**Figure 12**  
Woods Hole, Massachusetts, USA 2008 Aquadome



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
**Figure 11**  
Puerto Lindo, Panama 2008 Beta 2 Micropods



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
**Figure 13**  
South Korea, 2009 A3600

water. By similar logic, the small net pen, as a whole, because of its shorter stream-wide dimension experiences more water exchanges in a given current. From the perspective of an individual fish in an ocean net pen, water quality depends on how many other fish are metabolizing between it and its source of clean and oxygenated water from outside the containment. At constant  $\text{kg}/\text{m}^3$ , interior fish in a large net pen experience significantly degraded water quality compared to interior fish in a small net pen.

The potential advantage of low-volume, high-density (LVHD) net pen culture is not a new concept.

Schmittou pioneered the concept of LVHD in 1968, and the approach has since been applied in a project sponsored by the American Soybean Association in China [4]. Growing densities of up to  $200 \text{ kg}/\text{m}^3$  are recorded. McMaster et al. [5] suggests stocking density ranges for Florida pompano between 35 and  $100 \text{ kg}/\text{m}^3$ . Even at close to typical salmon stocking densities, the economics of small net pens become apparent. Table 1 shows a constant crop yield per capital investment in containment as density in the smaller net pens is increased.



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
Figure 14

South Korea, 2009 3-A3600



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
Figure 16

Mexico, 2009 A3600 under tow



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
Figure 15

Panama, 2010 A132 copper mesh



**Aquapod Systems for Sustainable Ocean Aquaculture.**  
Figure 17

Sorrento, ME USA A400





**Aquapod Systems for Sustainable Ocean Aquaculture.**

**Figure 18**

Mexico, 2010 2-A3600, 1-A212

In addition to this economic advantage, strong, secure small net pens will alleviate some of the economic barrier to entry for offshore fish farming. Small net pens require smaller vessels and support equipment; they are easily shipped, assembled, and deployed. In the near term, OFT sees a potential market for net pen hardware export to developing world economies that are transitioning from capture fisheries to fish farming, and who want to skip the near-shore environmental problems and user conflicts by moving directly offshore. This includes third world artisanal fishermen who in global aggregate contribute to marine fish resource depletion. Also, the mooring gear required for smaller net pens is lighter and more secure; scuba diving is reduced – especially deep diving; harvesting is easier; small pens are easier to tow; and risk is diversified.



**Aquapod Systems for Sustainable Ocean Aquaculture. Figure 19**

Mexico 2010 A212 under tow

**Aquapod Systems for Sustainable Ocean Aquaculture. Table 1** The 212 m<sup>3</sup> net pen stocked at 51 kg/m<sup>3</sup> gives interior fish better access to clean water (less biomass between center of net pen and net) and an equal yield per capital cost of containment

| Yield from various-sized Aquapod net pens |            |               |                                   |                                             |                                               |                    |                                                    |
|-------------------------------------------|------------|---------------|-----------------------------------|---------------------------------------------|-----------------------------------------------|--------------------|----------------------------------------------------|
| Size of pen m <sup>3</sup>                | Radius (m) | Diameter (ft) | Approximate cost of aquapod (USD) | Final stocking density (kg/m <sup>3</sup> ) | Biomass of fish between center of pen and net | Yield per pen (kg) | Yield (kg) per capital cost of containment (kg/\$) |
| 7,000                                     | 11.87      | 77.8          | \$218,000                         | 17                                          | 202                                           | 119,000            | 0.55                                               |
| 3,600                                     | 9.51       | 62.4          | \$140,000                         | 21                                          | 200                                           | 75,600             | 0.54                                               |
| 212                                       | 4.04       | 26.7          | \$20,000                          | 51                                          | 189                                           | 10,812             | 0.54                                               |

### Use of Small Aquapods for Artisanal Fish Farming

According to the FAO, “Artisanal Fisheries” are traditional fisheries involving fishing households (as opposed to commercial companies), using relatively small amount of capital and energy, relatively small fishing vessels or canoes, often beach-based, making short fishing trips, close to shore, mainly for local consumption. Artisanal fisheries can be subsistence or commercial fisheries, providing for local consumption or even export. They are sometimes referred to as small-scale fisheries or day fisheries.

Numbers of artisanal fishermen are hard to obtain because the sector is informal, often outside the regulated fishery (some illegal) and part-time. In Ghana, for example, canoe fishermen go to sea for about 150 days a year. Half the catch is obtained in 2 months, so it is obviously a part-time endeavor. The FAO estimates that as much as 95% of the annual catch in Africa is from artisanal fisheries. In Mexico, the FAO estimates that 96.6% of the fishing fleet (108,205 boats) are

*panga* (skiffs). Worldwide, the UN estimates there are 35-million people whose primary livelihood comes directly from fishing, and that artisanal fishermen outnumber large- and medium-scale fishermen by 4 or 5 to 1. This would put the number of artisanal fishermen in the range of 28 million to 30 million individuals worldwide (Table 2).

**From NIOSH Second Conference on International Fishing Industry Safety and Health** Overexploitation of coastal resources and decreasing catches has forced subsistence artisanal fishermen to go further offshore in search of fish in vessels designed for near-shore use. In a study done in 2000, the FAO reports an alarming increase in fishing-related fatalities as a result.

Small-scale marine aquaculture is one way to bring sustainability and consistency to this often-overlooked fishery sector. Turning artisanal fishermen into artisanal fish farmers will not be an easy task. However, small Aquapods have several advantages, beside

**Aquapod Systems for Sustainable Ocean Aquaculture. Table 2** Estimates of the number of fishing vessels in Pacifica [3]

| Location           | Locally based vessels [3] | Motorized artisanal fishing vessels [6] | Nonmotorized artisanal fishing vessels [6] |
|--------------------|---------------------------|-----------------------------------------|--------------------------------------------|
| Cook Islands       | 10 L/L                    | 200                                     | 120                                        |
| Fiji               | 96 L/L; 1 P/L             | 1,600                                   | 400                                        |
| Fed. States Micro. | 34 L/L; 8 P/S             | 2,000                                   | 600                                        |
| Kiribati           | 2 L/L; 1 P/S              | 600                                     | 5,000                                      |
| Marshall Islands   | 54 L/L; 5 P/S             | 500                                     | 250                                        |
| Nauru              | 1 L/L                     | 100                                     | 80                                         |
| Niue               | 100 skiffs                | 60                                      | 240                                        |
| Palua              | 71 L/L; 1 P/S             | 700                                     | 40                                         |
| PNG                | 40 L/L; 24 P/S            | 8,000                                   | 10,000                                     |
| Samoa              | 153 L/L                   | 80                                      | 100                                        |
| Solomon Islands    | 8 L/L; 2 P/S; 12 P/L      | 1,800                                   | 5,000                                      |
| Tonga              | 26 L/L                    | 800                                     | 200                                        |
| Tuvalu             | 20 skiffs                 | 200                                     | 500                                        |
| Vanatu             | 10 skiffs                 | 250                                     | 500                                        |
| Total              | 495 L/L; 40 P/S; 14 P/L   | 16,890                                  | 24,530                                     |

P/L pole and line vessel, P/S purse seiner, L/L longliner

affordability, that makes them advantageous for artisanal aquaculture:

- Size of boats and equipment can be smaller
  - The 7–10-m diameter Aquapods can be serviced with existing artisan vessels, whether motorized or not. Equipment needs are minimized and may include a small air compressor or pump.
- Ease of assembly and deployment – lighter mooring gear
  - Small Aquapods are easy to assemble, and can be put together by several men using hand tools on a beach and rolled into the water (or rolled out of the water for harvest). The small sectional area means a relatively low drag resistance, so mooring gear can be much lighter than traditional net pens.
- Single-point moorings are simple and inexpensive
  - An attractive option from a cost and environmental standpoint for these small pens is a single-point mooring, allowing the net pen to swing in a watch circle just as a boat moored in a harbor. Low drag forces on the anchor make this feasible.
- Less deep diving, most diving can be done by hookah
  - A net pen less than 10 m in diameter, when brought to the surface for maintenance, will allow almost all underwater activity to be performed without scuba gear. The spherical shape of the pen and the ability to rotate it allows for all portions of the net to be surfaced.
- More available deployment areas due to shallower water depth
  - Conventionally sized submersible pens require water depths of more than 30 m. The small Aquapods can be deployed in water depths of 12–25 m. This means boat trips far offshore are not necessary.
- Easier harvest – selective harvest by netting or whole pen harvest is possible.
- Diversification of risk as opposed to “all eggs in one basket.”
- Ease of towing – their small size and rigidity facilitate easy repositioning, or towing of a pen to market port.
- Predator proof

- Sharks, seals, and sea lions are a huge risk to artisanal fisheries – in some places, sharks are one of the main targets of artisanal fishermen. According to the Chamber of Commerce in Somalia, sharks constitute 90% of the artisanal catch. Artisanal fish farmers cannot tolerate this risk to the crop and to human safety, and the shark proof Aquapod is the solution.
- Storm resistant
  - The Aquapod net pens are submersible, so they can be submerged below storm waves when a hurricane or typhoon approaches, eliminating a potentially devastating risk.

### Economics of Artisanal Fish Farming with Aquapods

Catch estimates from artisanal fishermen are difficult to obtain – most are subsistence harvests, some excess is sold locally, and a little is exported. The sparse data that exists can only estimate the amount of fish that are consumed by the fisherman’s family and community based on how much is sold. However, FAO estimates of annual catches range from 2.5 t per year per boat (Yemen) to 3.6 t per year per (Mexico). Considering that the smallest Aquapod (212m<sup>3</sup>) could easily contain 5–10 t of fish at harvest, several of these small pens could probably produce more harvest than most artisanal fishermen ever see in a year.

Following is a spreadsheet analysis of what the economics of a small artisanal fish farm would look like. The assumptions include realistic and current pricing for commercial fish farm expenses, as low-volume pricing and sourcing of feed and fingerlings is unknown, and most likely would require some public sector support. As with all producing units, the sensitivity of market price trumps all other variables. Labor costs are assumed to be nil, as this is a family-scale operation.

The establishment of artisanal fish farms will likely be a private/public cooperative effort with governments and/or NGO’s subsidizing some capital costs, providing low interest loans, organizing distribution of feed and fingerlings, and providing training to artisanal fishermen as incentives to evolve from dependence on shrinking capture fisheries to a sustainable farming economy.

| Artisanal Aquapod Farm                             |           |                                |          |       |
|----------------------------------------------------|-----------|--------------------------------|----------|-------|
| A212 Aquapod Net Pens (No. of Aquapod Net Pens: 4) |           |                                |          |       |
| Sales                                              | \$212,000 | Assumptions                    |          |       |
| Cost of goods sold                                 |           | Biological FCR                 | 1.7      |       |
| Feed                                               | \$124,879 | Feed cost per kg               | \$1.65   |       |
| Fingerling                                         | \$34,980  | Harvest size in kg             | 2.0      |       |
| Outside services                                   | \$1,000   | Price of fingerlings           | \$1.50   | ea.   |
| Dive/Scuba expenses                                | \$2,000   | m <sup>3</sup> of growing vol. | 848      |       |
| Workboat expenses                                  | \$3,000   | Stocking final density         | 50       | kg/m3 |
| Fuel and oil expense                               | \$1,500   | Number of fish harvested       | 21,200   |       |
| Small tools                                        | \$250     | Kg Harvested                   | 42,400   |       |
| Concession                                         | \$500     | Sale price                     | \$5.00   | Kg    |
| Consumables                                        | \$750     | Amortization period            | 7        | years |
| Maintenance                                        | \$1,500   | Aquapod A212 net pen cost      | \$19,000 | ea.   |
| Miscellaneous                                      | \$800     | Mooring cost                   | \$3,000  | ea.   |
| Fish health                                        | \$1,000   | Total capital cost             | \$88,000 |       |
| Amortization of Aquapods                           | \$12,571  | Annual amortization expense    | \$12,571 |       |
| Cost of goods sold                                 | \$184,730 |                                |          |       |
| Gross margin                                       | \$27,270  |                                |          |       |
| Other overhead                                     | \$1,200   |                                |          |       |
| Net Income                                         | \$26,070  |                                |          |       |

## Bibliography

- DeCew J, Page S, Turmelle CA, Irish J (2006) Interior current reduction of the Aquapod™ cage. In: OCEANS 2006, OCEANS'06MTS/IEEE, Boston
- Caccese V (2006) Structural testing of various configurations for the aquapod net pen. Report No. C2004-018-001. Department of Mechanical Engineering, University of Maine, Maine, 38 pp

- Gillett R (2003) Aspects of sea safety in the fisheries of pacific island countries. FAO, Rome, 12
- Swick RA (2001) Feed based tilapia culture. Technical Bulletin AQ49-2001, American Soybean Association
- McMaster MF, Kloth TC, Coburn JF (2003) Prospects for commercial pompano mariculture. Aquaculture America 2003
- McCoy (1991) NIOSH Second Conference on International Fishing Industry Safety and Health 1991

## Atmosphere-Water Exchange

BERND JÄHNE

Institute for Environmental Physics and Heidelberg Collaboratory for Image Processing (HCI), University of Heidelberg, Heidelberg, Germany

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Mechanisms of Atmosphere–Water Exchange  
 Experimental Techniques and Results  
 Empiric Parametrization  
 Future Directions  
 Bibliography

## Glossary

**Bulk coefficients**  $c_i$  relate the *transfer velocity*  $k$  for a species  $i$  to the *wind velocity*  $U_r$  in a reference height, typically at 10 m above the mean water level:  $c_i = k_i/U_r$ . From the bulk coefficient, the flux density  $j_i$  of a species can be computed as  $j_i = c_i(C_r - C_0)U_r$ , where  $C_r$  and  $C_0$  are the corresponding concentrations at the reference height and right at the water surface, respectively. For momentum density ( $\rho U$ ) the bulk coefficients is also known as the *drag coefficient*  $c_D$ . It can also be expressed as  $c_D = (u_*/U_r)^2$  with the momentum flux given by  $j_m = \rho u_*^2$ ;  $u_*$  is the *friction velocity*.

**Friction velocity**  $u_*$  is a measure for the tangential force per area applied by the wind at the water surface, the shear stress  $\tau = \rho u_*^2$ , which is also equal to the vertical momentum flux density  $j_m$ .

**Mass boundary layer** Thickness of the layers at both sides of the water surface in which transport of mass by turbulence is smaller than by molecular diffusion.

**Schmidt and Prandtl numbers,  $Sc$  and  $Pr$**  The Schmidt and Prandtl numbers are the ratio of kinematic viscosity  $\nu$  (molecular diffusion coefficient for momentum) and the molecular diffusion coefficients for the corresponding chemical species,  $D$ , and heat,  $D_h$ , respectively. Thus, these numbers express how much slower chemical species and heat, respectively, are transported by molecular processes than momentum. In air, these numbers are in the order of one; in water, the Prandtl number is about 10 and the Schmidt number about 1,000.

**Transfer velocity  $k$**  is the velocity by which a momentum, heat, and chemical are transported across the surface; because of the concentration discontinuity at the water surface, the transfer velocity on the air side is different from the transfer velocity on the water side.

**Viscous boundary layer** Thickness of the layers at both sides of the water surface in which transport of momentum by turbulent is smaller than by molecular friction, resulting in a linear velocity profile in this layer.

## Definition of the Subject

Gaseous and volatile chemical species reside not only in the atmosphere. Because they dissolve in water, they are also distributed in the hydrosphere. The by far largest part of the hydrosphere is the ocean. Therefore, the exchange between atmosphere and oceans is the most important process for the fate of gaseous and volatile chemical species (Table 1).

For long-lived species, that is, life time by reaction larger than residence times in either the atmosphere or the ocean, two basic factors determine the fate of these species. For time scales longer than the turnover time of ocean mixing, which is much slower than the same process in the atmosphere, the solubility of the species determines where it ends up. On much shorter time scales, the speed by which the species are exchanged between atmosphere and ocean and the concentration difference determines the flux between these two compartments.

**Atmosphere-Water Exchange. Table 1** Comparison of size (as depth in m over entire earth surface) and mass of atmosphere and global water storage

| System                | Depth (m) | Mass ( $10^{18}$ kg) |
|-----------------------|-----------|----------------------|
| Oceans                | 2,624     | 1,382                |
| Ice caps and glaciers | 47        | 25                   |
| Total ground water    | 46        | 24                   |
| Lakes                 | 0.35      | 0.18                 |
| Rivers                | 0.004     | 0.002                |
| Atmosphere            | 8,000     | 5.3                  |

While turbulent mixing is quite fast within the ocean and the atmosphere, it becomes more and more inefficient, the closer a species is transported to the surface. The final step toward the interface must be overcome by molecular diffusion, which is a slow process, especially in water. Therefore, very thin mass boundary layers at both sides of the water surface, in which turbulent transport is slower than molecular transport, are the bottleneck of the exchange between atmosphere and ocean.

In this way, tiny layers only about 20–200  $\mu\text{m}$  thick in the water and 200–2,000  $\mu\text{m}$  thick in the air control the exchange process. In consequence, the crucial question is which processes determine the thickness of these layers.

## Introduction

Research in air–water gas exchange dates back more than a century. Bohr [1], for example, was one of the first to study the exchange of carbon dioxide in a stirred bottle. With the early observation of increasing carbon dioxide concentration in the atmosphere interest in air–sea gas exchange rose considerably, triggered by the question how much of the excess carbon dioxide would go into the ocean [2, 3]. Since then significant and continuous research work was devoted for a better understanding of air–water mass exchange, but the mechanisms of the exchange processes are still not fully understood even today.

This entry is divided into three major parts and partly based on previous reviews of the author on the same subject [4–6]. Section “[Mechanisms of Atmosphere–Water Exchange](#)” describes the basic



mechanisms of air–water exchange and the many parameters influencing it. Focusing on field data, section “[Experimental Techniques and Results](#)” summarizes the various experimental techniques and gives a synthesis of the available experimental data. Using the information collected in the previous two sections, section “[Empiric Parametrization](#)” tries to give the best possible empirical parametrization of the exchange rate and section “[Future Directions](#)” points to future directions of research.

## Mechanisms of Atmosphere–Water Exchange

### Turbulent and Molecular Transport

The transfer of gases and volatile chemical species between the atmosphere and oceans is driven by a concentration difference and the transport by molecular and turbulent motion. Both types of transport processes can be characterized by *diffusion coefficients*, denoted by  $D$  and  $K_c$ , respectively. The resulting flux area density  $j_c$  is proportional to the diffusion coefficient and the concentration gradient:

$$j_c = -(D + K_c(z)) \nabla_c. \quad (1)$$

In a stationary homogeneous case and without sinks and sources by chemical reactions, the flux density  $j$  is in vertical direction and constant. Then integration of [Eq. 1](#) yields vertical concentration profiles:

$$C(z_r) - C(0) = j_c \int_0^{z_r} \frac{1}{D + K_c(z)} dz. \quad (2)$$

The molecular diffusion coefficient is proportional to the velocity of the molecules and the free length between collisions. The same concept can be applied to turbulent diffusion coefficients. Far away from the interface, the free length (called *mixing length*) is set proportional to the distance from the interface and the turbulent diffusion coefficient  $K_c$  for mass transfer is:

$$K_c = \frac{\kappa}{Sc_t} u_* z, \quad (3)$$

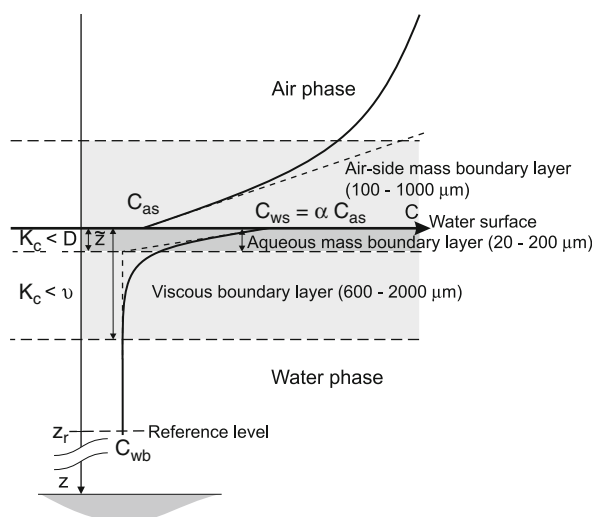
where  $\kappa = 0.41$  is the *von Kármán constant*,  $u_*$ , the *friction velocity*, a measure for the velocity fluctuations in a turbulent flow, and  $Sc_t = K_m/K_c$  the *turbulent Schmidt number*. Closer to the interface, the turbulent

diffusion coefficients are decreasing even faster. Once a critical length scale  $l$  is reached, the Reynolds number  $Re = u_* l/\nu$  ( $\nu$  is the kinematic viscosity, the molecular diffusion coefficient for momentum) becomes small enough so that turbulent motion is attenuated by viscosity. The degree of attenuation depends on the properties of the interface. At a smooth solid wall,  $K_c \propto z^3$ , at a free water interface it could be in the range between  $K_c \propto z^3$  and  $K_c \propto z^2$  depending on surface conditions.

### Viscous and Mass Boundary Layers

Boundary layers are formed on both sides of the interface ([Fig. 1](#)). When the turbulent diffusivity becomes equal to the kinematic viscosity, the edge of the *viscous boundary layer* is reached. As the name implies, this layer is dominated by viscous dissipation and the velocity profile becomes linear because of a constant diffusivity. The edge of the *mass boundary layer* is reached when the turbulent diffusivity becomes equal to the molecular diffusivity. The relative thickness of both boundary layers depends on the dimensionless ratio  $Sc = \nu/D$  (*Schmidt number*).

The viscous and mass boundary layers are of about the same thickness in the air, because values of  $D$  for



**Atmosphere-Water Exchange. Figure 1**  
Schematic graph of the mass boundary layers at a gas–liquid interface for a tracer with a solubility  $\alpha = 3$

various gaseous species and momentum are about the same ( $Sc_{air}$  is 0.56 for  $H_2O$ , 0.63 for heat, and 0.83 for  $CO_2$ ). In the liquid phase, the situation is completely different. With Schmidt numbers in the range from 100 to 3,000 (Fig. 2, Table 3), molecular diffusion for a dissolved volatile chemical species is two to three orders of magnitude slower than diffusion of momentum (Table 2). Thus, the mass boundary layer is significantly thinner than the viscous boundary layer in the liquid phase. This means that the transfer of gases is

much slower and almost entirely controlled by the tiny residual turbulence in the small top fraction of the viscous boundary layer. This basic fact makes it difficult to investigate the mechanism of air–water gas transfer both theoretically and experimentally. In addition, the transfer process depends strongly on the water temperature because the Schmidt number decreases by about a factor of 6 from 0°C to 35°C (Fig. 2, Table 3).

### Description of Transport

The amount of species exchanged between the air and water across the interface can be described by a quantity with the units of a velocity. It represents the velocity with which a tracer is pushed by an imaginary piston across the surface. This quantity is known as the *transfer velocity*  $k$  (also known as the *piston velocity*, *gas exchange rate* or *transfer coefficient*). It is defined as the flux density divided by the concentration difference between the surface and the bulk at some reference level  $z_r$ :

$$k = \frac{j_c}{C_s - C_b}. \quad (4)$$

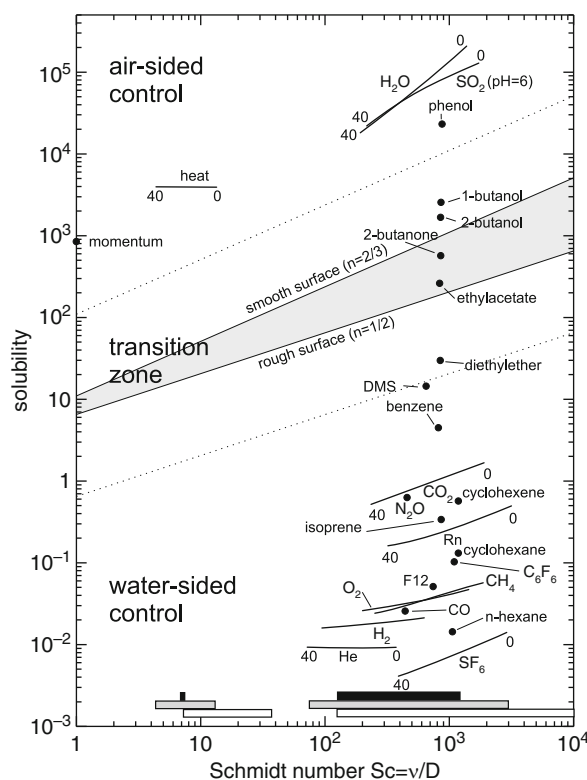
The inverse of the transfer velocity is known as the *transfer resistance*  $R$ :

$$R = \frac{C_s - C_b}{j_c}. \quad (5)$$

The indices  $s$  and  $b$  denote the surface and bulk. Both quantities can directly be related to vertical concentration profiles by introducing Eqs. 4 and 5 into Eq. 2:

$$R = \frac{1}{k} = \int_0^{z_r} \frac{1}{D + K_e(z)} dz. \quad (6)$$

Thus, the transfer resistances over several height intervals can be added in the same way as electrical resistances that are connected in series. Typical values of the transfer velocity across the water-side mass boundary layer are  $10^{-6}$ – $10^{-5}$  m/s (1–10 m/day). With respect to typical mixed layer depths in the ocean of about 100 m, gas transfer is a very slow process. It takes a time constant  $\tau = h/k = 10$ –100 days for the concentration of dissolved gases in the mixed layer to come into equilibrium with the atmosphere.



**Atmosphere-Water Exchange. Figure 2**

Schmidt number/solubility diagram including various volatile tracers, momentum, and heat for a temperature range (°C) as indicated. Filled circles refer to only a temperature of 20°C. The regions for air-side, mixed, and water-side control of the transfer process between gas and liquid phase are marked. At the solid lines, the transfer resistance is equal in both phases. The following dimensional transfer resistances were used:  $r_a = 31$ ,  $r_w = 12Sc^{2/3}$  (smooth),  $r_w = 6.55Sc^{1/2}$  (wavy surface) with  $r_a = R_a u_{*a}$  and  $r_w = R_w u_{*w}$  (Adapted from Jähne and Haußecker [6])

**Atmosphere-Water Exchange. Table 2** Diffusion coefficients for various gases and volatile chemical species in deionized water and in some cases in sea water collected from Jähne et al. [18] and King et al. [20]. The data for momentum (kinematic viscosity) are from Sündermann [22]. Column 3 and 4 contain the parameters for the fit of the diffusion coefficient:  $D = A \exp[-E_a/(RT)]$ , the last four columns the diffusion coefficients for 5°C, 15°C, 25°C, and 35°C

| Species                                                         | Molecular mass | $A$ ( $10^{-5}$ cm <sup>2</sup> /s) | $E_a$ (kJ/mole) | $\sigma$ (Fit) % | Diffusion coefficient ( $10^{-5}$ cm <sup>2</sup> /s) |        |        |        |
|-----------------------------------------------------------------|----------------|-------------------------------------|-----------------|------------------|-------------------------------------------------------|--------|--------|--------|
|                                                                 |                |                                     |                 |                  | 5°C                                                   | 15°C   | 25°C   | 35°C   |
| Momentum                                                        | –              |                                     |                 |                  | 1,517                                                 | 1,139  | 893.4  | 724.3  |
| Momentum <sup>a</sup>                                           | –              |                                     |                 |                  | 1,560                                                 | 1,182  | 934.5  | 763.9  |
| Heat                                                            | –              | 379.2                               | 2.375           |                  | 135.80                                                | 140.72 | 145.48 | 150.08 |
| <sup>3</sup> He <sup>b,c</sup>                                  | 3.02           | 941                                 | 11.70           | 2.1              | 5.97                                                  | 7.12   | 8.39   | 9.77   |
| <sup>4</sup> He                                                 | 4.00           | 818                                 | 11.70           | 2.1              | 5.10                                                  | 6.30   | 7.22   | 8.48   |
| <sup>4</sup> He <sup>a</sup>                                    |                | 886                                 | 12.02           | 1.8              | 4.86                                                  | 5.88   | 7.02   | 8.03   |
| Ne                                                              | 20.18          | 1,608                               | 14.84           | 3.5              | 2.61                                                  | 3.28   | 4.16   | 4.82   |
| Kr                                                              | 83.80          | 6,393                               | 20.20           | 1.6              | 1.02                                                  | 1.41   | 1.84   | 2.40   |
| Xe                                                              | 131.30         | 9,007                               | 21.61           | 3.5              | 0.77                                                  | 1.12   | 1.47   | 1.94   |
| <sup>222</sup> Rn <sup>b</sup>                                  | 222.00         | 15,877                              | 23.26           | 11               | 0.68                                                  | 0.96   | 1.34   | 1.81   |
| H <sub>2</sub>                                                  | 2.02           | 3,338                               | 16.06           | 1.6              | 3.17                                                  | 4.10   | 5.13   | 6.23   |
| H <sub>2</sub> <sup>a</sup>                                     |                | 1,981                               | 14.93           | 4.3              | 3.05                                                  | 3.97   | 4.91   | 5.70   |
| CH <sub>4</sub>                                                 | 16.04          | 3,047                               | 18.36           | 2.7              | 1.12                                                  | 1.48   | 1.84   | 2.43   |
| CO <sub>2</sub>                                                 | 44.01          | 5,019                               | 19.51           | 1.3              | 1.07                                                  | 1.45   | 1.91   | 2.43   |
| DMS <sup>b</sup>                                                | 62.13          | 2,000                               | 18.10           |                  | 0.80                                                  | 1.05   | 1.35   | 1.71   |
| CH <sub>3</sub> Br <sup>b</sup>                                 | 94.94          | 3,800                               | 19.10           |                  | 0.98                                                  | 1.31   | 1.71   | 2.20   |
| F <sub>12</sub> <sup>b</sup> (CCl <sub>2</sub> F <sub>2</sub> ) | 120.91         | 4,100                               | 20.50           |                  | 0.58                                                  | 0.79   | 1.05   | 1.37   |
| F <sub>11</sub> <sup>b</sup> (CCl <sub>3</sub> F)               | 137.37         | 3,400                               | 20.00           |                  | 0.60                                                  | 0.81   | 1.07   | 1.38   |
| SF <sub>6</sub> <sup>b</sup>                                    | 146.05         | 2,900                               | 19.30           |                  | 0.69                                                  | 0.92   | 1.20   | 1.55   |

<sup>a</sup>In sea water

<sup>b</sup>Values of diffusion coefficients from fit, not measured values

<sup>c</sup>Set 15% higher than <sup>a</sup>He

**Boundary Layer Thickness** The *boundary layer thickness*  $\tilde{z}$  is defined as the thickness of a fictional layer in which the flux is maintained only by molecular transport:  $j = D(C_s - C_b)/\tilde{z}$ . Then with Eq. 4 the boundary layer thickness is given by:

$$\tilde{z} = \frac{D}{k}. \quad (7)$$

Geometrically,  $\tilde{z}$  is given as the intercept of the tangent to the concentration profile at the surface and the bulk (Fig. 1). With thicknesses between 20 and 200  $\mu\text{m}$ , the mass boundary layer is extremely thin.

**Boundary Layer Time Constant** The *time constant*  $\tilde{t}$  for the transport across the mass boundary layer is given by:

$$\tilde{t} = \frac{\tilde{z}}{k} = \frac{D}{k^2}. \quad (8)$$

Typical values for  $\tilde{t}$  are 0.04–4 s. Any chemical reaction with a time constant larger than  $\tilde{t}$  does not significantly affect the transfer process. Therefore, CO<sub>2</sub> can be regarded as an inert gas, but not fast hydrating acid gases such as SO<sub>2</sub>.

**Atmosphere-Water Exchange. Table 3** Schmidt numbers of various gases and volatile species in the temperature range from 0°C to 40°C

A

| Species                                            | Schmidt number |       |       |       |       |      |      |      |      |
|----------------------------------------------------|----------------|-------|-------|-------|-------|------|------|------|------|
|                                                    | 0°C            | 5°C   | 10°C  | 15°C  | 20°C  | 25°C | 30°C | 35°C | 40°C |
| Heat                                               | 13.45          | 11.19 | 9.46  | 8.10  | 7.02  | 6.14 | 5.42 | 4.82 | 4.32 |
| <sup>3</sup> He                                    | 329            | 254   | 200   | 160   | 130   | 107  | 88   | 74   | 63   |
| <sup>4</sup> He                                    | 379            | 293   | 230   | 184   | 149   | 122  | 102  | 85   | 72   |
| <sup>4</sup> He <sup>a</sup>                       | 411            | 319   | 252   | 201   | 164   | 135  | 112  | 94   | 79   |
| Ne                                                 | 768            | 579   | 445   | 347   | 276   | 221  | 180  | 148  | 122  |
| Kr                                                 | 2,045          | 1,478 | 1,090 | 819   | 625   | 483  | 379  | 301  | 241  |
| Xe                                                 | 2,701          | 1,930 | 1,409 | 1,047 | 791   | 606  | 471  | 370  | 294  |
| <sup>222</sup> Rn                                  | 3,168          | 2,235 | 1,611 | 1,182 | 883   | 669  | 514  | 400  | 314  |
| H <sub>2</sub>                                     | 633            | 473   | 360   | 278   | 219   | 174  | 140  | 114  | 94   |
| H <sub>2</sub> <sup>a</sup>                        | 663            | 502   | 387   | 304   | 242   | 195  | 159  | 131  | 109  |
| CH <sub>4</sub>                                    | 1,908          | 1,400 | 1,047 | 797   | 616   | 483  | 383  | 308  | 250  |
| CO <sub>2</sub>                                    | 1,922          | 1,397 | 1,036 | 782   | 600   | 466  | 367  | 293  | 236  |
| DMS                                                | 2,593          | 1,905 | 1,428 | 1,089 | 844   | 662  | 527  | 423  | 344  |
| CH <sub>3</sub> Br                                 | 2,120          | 1,545 | 1,150 | 870   | 669   | 522  | 412  | 329  | 266  |
| F <sub>12</sub> (CCl <sub>2</sub> F <sub>2</sub> ) | 3,639          | 2,624 | 1,931 | 1,447 | 1,102 | 851  | 666  | 527  | 422  |
| F <sub>11</sub> (CCl <sub>3</sub> F)               | 3,521          | 2,549 | 1,883 | 1,416 | 1,082 | 839  | 658  | 523  | 420  |
| SF <sub>6</sub>                                    | 3,033          | 2,208 | 1,640 | 1,239 | 952   | 741  | 585  | 467  | 376  |

<sup>a</sup>In sea water

The definitions of the three parameters  $k$ ,  $\tilde{z}$ ,  $\tilde{t}$  are generally valid and do not depend on any models of the boundary layer turbulence. According to Eqs. 7 and 8, they are coupled via the molecular diffusion coefficient. Therefore, only one of them needs to be measured to get knowledge of all three parameters provided the molecular diffusion coefficient of the species is known.

### Partitioning of Transfer Between Air and Water

Because a mass boundary layer exists on both sides, it is important to determine which one controls the transfer, that is, exhibits the largest transfer resistance (or lowest transfer velocity). At the surface itself, the thermodynamic solubility equilibrium is assumed to be established between the tracer concentrations  $c_a$  in the gas phase and  $c_w$  in the liquid phase:

$$C_{ws} = \alpha C_{as}, \quad (9)$$

where  $\alpha$  is the dimensionless solubility (partition coefficient). A solubility  $\alpha \neq 1$  causes a concentration jump at the surface (Fig. 1). Thus, the resulting total transfer velocity  $k_t$  or transfer resistance  $R_t$  can either be viewed from the gas phase or the liquid phase. Adding them up, the factor  $\alpha$  must be considered to conserve the continuity of the concentration profile:

$$\text{air side: } \frac{1}{k_{at}} = \frac{1}{k_a} + \frac{1}{\alpha k_w}, \quad R_{at} = R_a + R_w/\alpha, \quad (10)$$

$$\text{water side: } \frac{1}{k_{wt}} = \frac{\alpha}{k_a} + \frac{1}{k_w}, \quad R_{wt} = \alpha R_a + R_w.$$

The total transfer velocities in air and water differ by the factor  $\alpha$ :  $k_{at} = \alpha k_{wt}$ . The ratio  $\alpha k_w/k_a$  determines which boundary layer controls the transfer process. A high solubility shifts control of the transfer process

to the gas-phase boundary layer, and a low solubility to the aqueous layer. The solubility value for a transition from air-sided to water-sided control depends on the ratio of the transfer velocities. Typically  $k_w$  is about 100–1,000 times smaller than  $k_a$ . Thus, the transfer of even moderately soluble volatile chemical species with solubilities up to 30 is controlled by the water side. Some environmentally important species lie in a transition zone where it is required to consider both transport processes (Fig. 2). The transfer of highly soluble volatile and/or chemically reactive gas is controlled by the air-side transfer process and thus analogous to the transfer of water vapor. The following considerations concentrate on the water-side transfer process.

### Gas Exchange at Smooth Water Surfaces

At smooth water surfaces, the theory of mass transfer is well established because it is equivalent to mass transfer to a smooth solid wall. The turbulent diffusivity can be described by the classic approach of Reichardt [7] with an initial  $z^3$  increase that smoothly changes to a linear increase in the turbulent layer as in Eq. 3. Then integration of Eq. 6 yields the following approximation for Schmidt numbers higher than 100 [12, 16]:

$$k_w = u_{*w} \frac{1}{12.2} Sc^{-2/3} \quad Sc > 60. \quad (11)$$

This equation establishes the basic analogy between momentum transfer and gas exchange. The transfer coefficients are proportional to the friction velocity in water, which describes the shear stress (tangential force per unit area)  $\tau = \rho_w u_{*w}^2$  applied by the wind field at the water surface. Assuming stress continuity at the water surface, the friction velocity in water is related to the friction velocity in air by:

$$u_{*w} = u_{*a} \left( \frac{\rho_a}{\rho_w} \right)^{1/2}. \quad (12)$$

The friction velocity in air,  $u_{*a}$ , can further be linked via the drag coefficient to the wind speed  $U_R$  at a reference height:  $c_D = (u_{*a}/U_R)^2$ . Depending on the roughness of the sea surface, the drag coefficient has values between 0.8 and  $2.4 \cdot 10^{-3}$ . In this way, the gas exchange rate is directly linked to the wind speed. The

gas exchange further depends on the chemical species and the water temperature via the Schmidt number.

### Influence of Waves

A free water surface is neither solid nor is it smooth as soon as short wind waves are generated. On a free water surface, velocity fluctuations are possible. Thus, there can be convergence or divergence zone at the surface; surface elements may be dilated or contracted. At a clean water surface, dilation or contraction of a surface element does not cause restoring forces, because surface tension only tries to minimize the total free surface area, which is not changed by this process. As a consequence of this hydrodynamic boundary condition, the turbulent diffusivity normal to the interface can now increase with the distance squared from the interface,  $K_c \propto z^2$ . Then:

$$k_w = u_{*w} \frac{1}{\beta} Sc^{-1/2}, \quad (13)$$

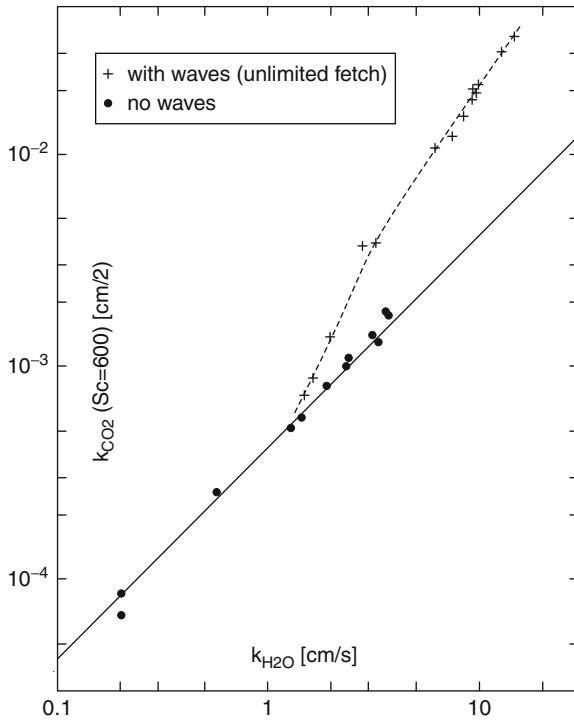
where  $\beta$  is a dimensionless constant.

In comparison to the smooth case in Eq. 11, the exponent  $n$  of the Schmidt number drops from  $-2/3$  to  $-1/2$ . This increases the transfer velocity for a Schmidt number of 600 by about a factor of 3. The total enhancement depends on the value of the constant  $\beta$ .

Wind waves cannot be regarded as static roughness elements for the liquid flow because their characteristic particle velocity is of the same order of magnitude as the velocity in the shear layer at the surface. This fact causes a basic asymmetry between the turbulent processes on the air and on the water sides of the interface. Therefore, the wave effect on the turbulent transfer in the water is much stronger and of quite different character than in the air. This basic asymmetry can be seen if the transfer velocity for  $\text{CO}_2$  is plotted against the transfer velocity for water vapor (Fig. 3, [15]). At a smooth water surface, the points fall well on the theoretical curve predicted by the theory for a smooth rigid wall. However, as soon as waves occur at the water surface, the transfer velocity of  $\text{CO}_2$  increases significantly beyond the predictions.

Even at high wind speeds, the observed surface increase is well below 20% [24]. When waves are





**Atmosphere-Water Exchange. Figure 3**

Transfer velocity of carbon dioxide plotted against the transfer velocity of water vapor. Measurements from a small annular wind/wave facility Jähne [15]

generated by wind, energy is not only transferred via shear stress into the water but a second energy cycle is established. The energy put by the turbulent wind into the wave field is transferred to other wave numbers by nonlinear wave-wave interaction and finally dissipated by wave breaking, viscous dissipation, and turbulence. The turbulent wave dissipation term is the least known term and of most importance for enhanced near-surface turbulence. Evidence for enhanced turbulence levels below wind waves has been reported from field and laboratory measurements. Experimental results also suggest that the gas transfer rate is better correlated with the *mean square slope* of the waves as an integral measure for the nonlinearity of the wind wave field than with the wind speed.

It is not yet clear, however, to what extent *microscale wave breaking* can account for the observed enhanced gas transfer rates. A gravity wave becomes unstable and generates a steep train of capillary waves at its leeward

face and has a turbulent wake. This phenomenon can be observed even at low wind speeds, as soon as wind waves are generated. At higher wind speeds, the frequency of microscale wave breaking increases.

At high wind speeds, wave breaking with the entrainment of bubbles may enhance gas transfer further. This phenomenon complicates the gas exchange between atmosphere and the oceans considerably [19, 26]. First, bubbles constitute an additional exchange surface. This surface is, however, only effective for gases with low solubility. For gases with high solubility, the gas bubbles quickly come into equilibrium so that a bubble takes place in the exchange only for a fraction of its life time. Thus bubble-mediated gas exchange depends – in contrast to the exchange at the free surface – on the solubility of the gas tracer. Second, bubble-mediated gas transfer shifts the equilibrium value to slight supersaturation due to the enhanced pressure in the bubbles by surface tension and hydrostatic pressure. Third, breaking waves also enhance near-surface turbulence during the breaking event and the resurfacing of submerged bubbles [17].

Experimental data are still too sparse for the size and depth distribution of bubbles and the flux of the bubbles through the interface under various sea states for a sufficiently accurate modeling of bubble-mediated air-sea gas transfer and thus a reliable estimate of the contribution of bubbles to the total gas transfer rate. Some experiments from wind/wave tunnels and the field suggest that significant enhancements can occur, other experiments could not observe a significant influence of bubbles.

### Influence of Surface Films

A film on the water surface creates pressure that works against the contraction of surface elements. This is the point at which the physicochemical structure of the surface influences the structure of the near-surface turbulence as well as the generation of waves [13, 21]. As at a rigid wall, a strong film pressure at the surface maintains a two-dimensional continuity at the interface just as at a rigid wall. Therefore, Eq. 11 should be valid for a smooth film-covered water surface and has indeed been verified in wind/wave tunnel studies as the lower limit for the transfer velocity. As a consequence,

both Eqs. 11 and 13 can only be regarded as limiting cases. A more general approach is required that has not yet been established. One possibility is a generalization of Eqs. 11 and 13 to:

$$k_w = u_{*w} \frac{1}{\beta(s)} Sc^{-n(s)}, \quad (14)$$

where both  $\beta$  and  $n$  depend on dimensionless parameters describing the surface conditions  $s$ . Even films with low film pressure may easily decrease the gas transfer rate to half of its value at clean water surface conditions. But still too few measurements at sea are available to establish the influence of surfactants on gas transfer for oceanic conditions more quantitatively.

## Experimental Techniques and Results

### Laboratory Facilities

Laboratory facilities play an important role in the investigation of air-sea gas transfer. Only laboratory studies allow a systematic study of the mechanisms and are thus an indispensable complement to field experiments. Almost all basic knowledge about gas transfer has been gained by laboratory experiments in the past.

### Geochemical Tracer Techniques

The first oceanic gas exchange measurements were performed using geochemical tracer methods such as the  $^{14}\text{C}$  [2],  $^3\text{He}/\text{T}$  [23], or  $^{222}\text{Rn}/^{226}\text{Ra}$  [8, 9] methods. The volume and time-average flux density is given by mass balance of the tracer concentration in a volume of water  $V_w$ :

$$V_w \dot{c}_w = F_w j \quad \text{or} \quad j = h_w \dot{c}_w, \quad (15)$$

where  $F_w$  and  $h_w$  are the surface area and the effective height  $V_w/F_w$  of a well-mixed water body, respectively. The time constant  $\tau_w = h_w/k$  is in the order of days to weeks. It is evident that the transfer velocities obtained in this way provide only values integrated over a large horizontal length scales and time scales in the order of  $\tau_w$ . Thus a parametrization of the transfer velocity is only possible under steady state conditions over extended periods. Moreover, the mass balance contains many other sources and sinks besides air-sea gas

exchange and thus may cause severe systematic errors in the estimation of the transfer velocity. Consequently, mass balance methods are only poorly suited for the study of the mechanisms of air–water gas transfer.

### Tracer Injection

The pioneering lake studies for tracer injection used sulfur hexafluoride ( $\text{SF}_6$ ). However, the tracer concentration decreases not only by gas exchange across the interface but also by horizontal dispersion of the tracer. This problem can be overcome by the *dual tracer technique* [10] simultaneously releasing two tracers with different diffusivities (e.g.,  $\text{SF}_6$  and  $^3\text{He}$ ). When the ratio of the gas transfer velocities of the two tracers is known, the dilution effect by tracer dispersion can be corrected, making it possible to derive gas transfer velocities. But the basic problem of mass balance techniques, that is, their low temporal resolution, remains also with artificial tracer approaches.

### Eddy Correlation Flux Measurements

Eddy correlation techniques are used on a routine basis in micrometeorology, that is, for tracers controlled by the boundary layer in air (momentum, heat, and water vapor fluxes). Direct measurements of the air-sea fluxes of gas tracers are very attractive because the flux densities are measured directly and have a much better temporal resolution than the mass balance based techniques. Unfortunately, large experimental difficulties arise when this technique is applied to gas tracers controlled by the aqueous boundary layer [Broecker 1986, 14]. The concentration difference in the air is only a small fraction of the concentration difference across the aqueous mass boundary layer. But after more than 20 years of research has this technique delivered useful results. Some successful measurements under favorable conditions have been reported and it appears that remaining problems can be overcome in the near future.

### Passive and Active Thermography

The basic idea of this technique is to determine the concentration difference across the mass boundary layer when the flux density  $j$  of the tracer across the

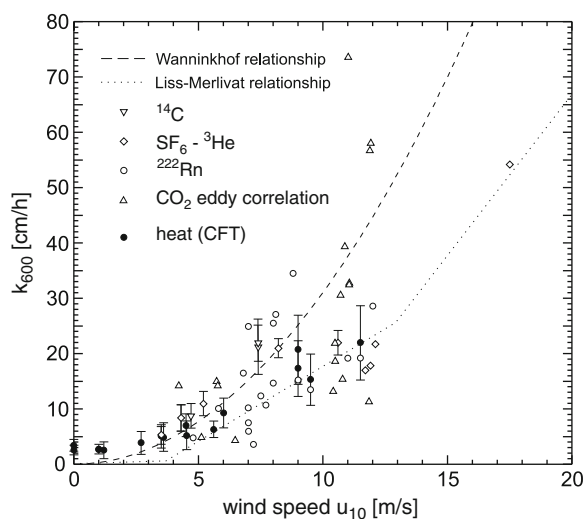
interface is known. The local transfer velocity can be determined by simply measuring the concentration difference  $\Delta c$  across the aqueous boundary layer (cold surface skin temperature) according to Eq. 4 with a time constant  $\tilde{t}$  for the transport across the boundary layer Eq. 8. This technique is known as the *controlled flux technique* (CFT).

Heat proves to be an ideal tracer for the CFT. The temperature at the water surface can then be measured with high spatial and temporal resolution using IR thermography. A known and controllable flux density can be applied by using infrared radiation. Infrared radiation is absorbed in the first few ten micrometers at the water surface. Thus a heat source is put right at top of the aqueous viscous boundary layer. Then the CFT directly measures the water-side heat transfer velocity.

A disadvantage of the CFT is that the transfer velocity of gases must be extrapolated from the transfer velocity of heat. The large difference in the Schmidt number (7 for heat, 600 for  $\text{CO}_2$ ) casts some doubt whether the extrapolation to so much higher Schmidt numbers is valid.

Two variants of the technique proved to be successful. Active thermography uses a  $\text{CO}_2$  laser to heat a spot of several centimeter in diameter on the water surface. The heat transfer rates are estimated from the temporal decay of the heated spot. Passive thermography uses the naturally occurring heat fluxes caused by latent heat flux  $j_b$ , sensible heat flux  $j_s$  and longwave emission of radiation  $j_r$ . The net heat flux  $j_n = j_l + j_s + j_r$  results according to Eq. 4 in a temperature difference across the interface of  $\Delta T = j_n / (\rho c_p k_h)$ . Because of the turbulent nature of the exchange process any mean temperature difference is associated with surface temperature fluctuations which can be observed in thermal images. With this technique, the horizontal structure of the boundary layer turbulence can be observed. Surface renewal is directly observable in the IR image sequences, which show patches of fluid being drawn away from the surface.

With some knowledge about the statistics of the temperature fluctuations, the *temperature difference*  $\Delta T$  across the interface as well as the time constant  $\tilde{t}$  of heat transfer can be computed from the temperature distribution at the surface. Results obtained with this technique are shown in Fig. 4.



**Atmosphere-Water Exchange. Figure 4**

Summary of gas exchange field data normalized to a Schmidt number of 600 and plotted versus wind speed together with the empirical relationships of Liss and Merlivat [11] and ? (After Jähne and Haußecker [6])

### Summary of Field Data

A collection of field data is shown in Fig. 3. Although the data show a clear increase of the transfer velocity with wind speed, there is substantial scatter in the data that can only partly be attributed to uncertainties and systematic errors in the measurements. Thus, in addition, the field measurements reflect the fact that the gas transfer velocity is not simply a function of the wind speed but depends significantly on other parameters influencing near-surface turbulence, such as the wind-wave field and the viscoelastic properties of the surface film.

### Empiric Parametrization

Given the lack of knowledge, all theories about the enhancement of gas transfer by waves are rather speculative and are not yet useful for practical application. Thus, it is still state of the art to use semiempiric or empiric parameterizations of the gas exchange rate with the wind speed. Most widely used is the parametrization of Liss and Merlivat [11]. It identifies three physically well-defined regimes (smooth, wave-influenced, bubble-influenced) and proposes

a piece-wise linear relation between the wind speed  $U$  and the transfer velocity  $k$ :

$$k = 10^{-6} \begin{cases} 0.472U(Sc/600)^{-2/3} & U \leq 3.6 \text{ m/s} \\ 7.971(U - 3.39)(Sc/600)^{-1.2} & U > 3.6 \text{ m/s and} \\ & U \leq 13 \text{ m/s} \\ 16.39(U - 8.36)(Sc/600)^{-1/2} & U > 13 \text{ m/s} \end{cases} \quad (16)$$

At the transition between the smooth and wavy regime, a sudden artificial jump in the Schmidt number exponent  $n$  from  $2/3$  to  $1/2$  occurs. This actually causes a discontinuity in the transfer rate for Schmidt number unequal to 600.

The empiric parametrization of Wanninkhof [25] simply assumes a quadratic increase of the gas transfer rate with the wind speed:

$$k = 0.861 \cdot 10^{-6} (s/m) U^2 (Sc/600)^{-1/2}. \quad (17)$$

Thus, this model has a constant Schmidt number exponent  $n = 1/2$ . The two parameterizations differ significantly (see Fig. 3). The Wanninkhof parametrization predicts significantly higher values. The discrepancy between the two parameterizations and many others proposed (up to a factor of two) mirrors the current uncertainty in estimating the air-sea gas transfer rate.

## Future Directions

In the past, progress toward a better understanding of the mechanisms of air–water gas exchange was hindered by inadequate measuring technology. However, new techniques have become available and will continue to become available that will give a direct insight into the mechanisms under both laboratory and field conditions. This progress will be achieved by interdisciplinary research integrating such different research areas as oceanography, micrometeorology, hydrodynamics, physical chemistry, applied optics, and image processing.

Optical and image processing techniques will play a key role because only imaging techniques give direct insight to the processes in the viscous, heat, and mass boundary layers on both sides of the air–water interface. Eventually, all key parameters including flow fields, concentration fields, and waves will be captured

by imaging techniques with sufficient spatial and temporal resolution. The experimental data gained with such techniques will stimulate new theoretical and modeling approaches.

## Bibliography

### Primary Literature

- Bohr C (1899) Definition und methode zur bestimmung der invasions- und evasionscoefficienten bei der auflösung von gasen in flüssigkeiten. Werthe der genannten constanten sowie der absorptionscoefficienten der kohlenensäure bei auflösung in wasser und in chlornatriumlösungen. *Ann Phys Chem* 68:500–525
- Bolin B (1960) On the exchange of carbon dioxide between the atmosphere and the sea. *Tellus* 12(3):274–281. ISSN 2153-3490
- Revelle R, Suess HE (1957) Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric CO<sub>2</sub> during the past decades. *Tellus* 9:18–27
- Jähne B (1982) Dry deposition of gases over water surfaces (gas exchange). In: Flothmann D (ed) Exchange of air pollutants at the air-earth interface (dry deposition). Battelle Institute, Frankfurt
- Jähne B (2009) Air-sea gas exchange. In: Steele JH, Turekian KK, Thorpe SA (eds) Encyclopedia ocean sciences. Elsevier, Boston, pp 3434–3444
- Jähne B, Haußecker H (1998) Air-water gas exchange. *Annu Rev Fluid Mech* 30:443–468
- Reichardt H (1951) Vollständige darstellung der turbulenten geschwindigkeitsverteilung in glatten leitungen. *Z Angew Math Mech* 31:208–219
- Peng TH, Broecker WS, Mathieu GG, Li Y-H, Bainbridge A (1979) Radon evasion rates in the Atlantic and Pacific oceans as determined during the geosecs program. *J Geophys Res* 84(C5):2471–2487
- Roether W, Kromer B (1984) Optimum application of the radon deficit method to obtain air–sea gas exchange rates. In: Brutsaert W, Jirka GH (eds) Gas transfer at water surfaces. Reidel, Hingham, pp 447–457
- Watson AJ, Upstill-Goddard RC, Liss PS (1991) Air-sea exchange in rough and stormy seas measured by a dual tracer technique. *Nature* 349(6305):145–147
- Liss PS, Merlivat L (1986) Air-sea gas exchange rates: introduction and synthesis. In: Buat-Menard P (ed) The role of air-sea exchange in geochemical cycling. Reidel, Boston, pp 113–129
- Deacon EL (1977) Gas transfer to and across an air-water interface. *Tellus* 29:363–374
- Frew NM (1997) The role of organic films in air-sea gas exchange. In: Liss PS, Duce RA (eds) The sea surface and global change. Cambridge University Press, Cambridge, pp 121–171
- Jacobs C, Nightingale P, Upstill-Goddard R, Kjeld JF, Larsen S, Oost W (2002) Comparison of the deliberate tracer method

- and eddy covariance measurements to determine the air/sea transfer velocity of CO<sub>2</sub>. In: Saltzman E, Donelan M, Drennan W, Wanninkhof R (eds) Gas transfer at water surfaces. Geophysical Monograph, vol 127. American Geophysical Union
15. Jähne B (1980) Zur Parametrisierung des Gasaustauschs mit Hilfe von Laborexperimenten. Dissertation, Institut für Umweltphysik, Fakultät für Physik und Astronomie, Univ. Heidelberg, <http://d-nb.info/810123614>. IUP D-145
  16. Jähne B (1987) Image sequence analysis of complex physical objects: nonlinear small scale water surface waves. In: Proceedings of 1st international conference on computer vision, London, pp 191–200
  17. Jähne B (1991) New experimental results on the parameters influencing air-sea gas exchange. In: Wilhelms SC, Gulliver JS (eds) Air-water mass transfer, Selected papers from the 2nd international symposium on gas transfer at water surfaces, Minneapolis, 11–14 Sep 1990. ASCE, pp 582–592
  18. Jähne B, Heinz G, Dietrich W (1987) Measurement of the diffusion coefficients of sparingly soluble gases in water. *J Geophys Res* 92(C10):10,767–10,776
  19. Keeling RF (1993) On the role of large bubbles in air-sea gas exchange and supersaturation in the ocean. *J Marine Res* 51:237–271
  20. King DB, Bryun WJD, Zheng M, Saltzman ES (1995) Uncertainties in the molecular diffusion coefficient of gases in water for use in the estimation of air-sea exchange. In: Jähne B, Monahan E (eds) Air-water gas transfer, Selected papers, 3rd international symposium on air-water gas transfer. AEON, Hanau, pp 13–22
  21. McKenna SP, Bock EJ (2006) Physicochemical effects of the marine microlayer on air-sea gas transport. In: Gade M, Hühnerfuss H, Korenowski GM (eds) Marine surface films: chemical characteristics, influence on air-sea interactions, and remote sensing. Springer, Berlin/Heidelberg, pp 77–91
  22. Sündermann J (ed) (1986) Landolt-Börnstein, vol V 3c, Oceanography. Springer, Heidelberg
  23. Torgersen T, Top Z, Clarke WB, Jenkins WJ (1977) A new method for physical limnology-tritium-helium-3 ages results for Lakes Erie, Huron, and Ontario. *Limnol Oceanogr* 22:181–193
  24. Tschiersch J, Jähne B (1980) Gas Exchange through a rough water surface in a circular windtunnel; Wave characteristics under limited and unlimited fetch. In: Broecker HC, Hasse L (eds) Berichte aus dem Sonderforschungsbereich 94 Meeresforschung – Symposium on capillary waves and gas exchange, Trier, 2–6 July 1979, number 17. Univ. Hamburg, pp 63–70
  25. Wanninkhof R (1992) Relationship between wind speed and gas exchange over the ocean. *J Geophys Res* 97: 7373–7382
  26. Woolf D, Leifer I, Nightingale P, Rhee T, Bowyer P, Caulliez G, de Leeuw G, Larsen S, Liddicoat M, Baker J, Andreae M (2007) Modelling of bubble-mediated gas transfer: fundamental principles and a laboratory test. *J Marine Syst* 66:71–91

## Books and Reviews

- Banner ML (ed) (1999) The wind-driven air-sea interface, electromagnetic and acoustic sensing, wave dynamics and turbulent fluxes, proceedings of the symposium Sydney, Australia, 11–15 January 1999. University of New South Wales, Sydney
- Bengtsson LO, Hammer CU (eds) (2001) Geosphere-biosphere interactions and climate. Cambridge University Press, Cambridge
- Borges AV, Wanninkhof R (2007) Fifth international symposium on gas transfer at water surfaces. *J Mar Syst* 66(1–4):1–308
- Brasseur GP, Prinn RG, Pszenny AA (eds) (2003) Atmospheric chemistry in a changing world. An integration and synthesis of a decade of tropospheric chemistry research. Springer, Berlin
- Brutsaert W, Jirka GH (eds) (1984) Gas transfer at water surfaces. Reidel, Dordrecht
- Csanady GT (2001) Air-sea interaction, laws and mechanisms. Cambridge University Press, Cambridge
- Cussler EL (2009) Diffusion – mass transfer in fluid systems, 3rd edn. Cambridge University Press, Cambridge
- Danckwerts PV (1970) Gas-liquid reactions. MacGraw-Hill, New York
- Davies JT (1972) Turbulence phenomena. An introduction to the eddy transfer of momentum, mass, and heat, particularly at interfaces. Academic, New York/London
- Davies JT, Rideal EK (1963) Interfacial phenomena, 2nd edn. Academic, New York
- Dobson F, Hasse L, Davis R (eds) (1980) Air-sea interaction: instruments and methods. Plenum, New York
- Donelan MA, Hui WH, Plant WJ (eds) (1996) The air-sea interface, radio and acoustic sensing, turbulence and wave dynamics, proceedings of the symposium, Marseille, France, 24–30 June 1993. Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida
- Donelan MA, Drennan WM, Saltzman ES, Wanninkhof R (eds) (2002) Gas transfer at water surfaces. American Geophysical Union, Washington, DC
- Emerson SR, Hedges J (2008) Chemical oceanography and the marine carbon cycle. Cambridge University Press, Cambridge
- Favre A, Hasselmann K (eds) (1978) Turbulent fluxes through the sea surface, wave dynamics, and prediction, proceedings of the symposium, Marseille, 1977. Plenum, New York
- Fogg PGT, Sangster J (2003) Chemicals in the atmosphere: solubility, sources, and reactivity. Wiley, Chichester
- Gade M, Hühnerfuss H, Korenowski GM (eds) (2005) Marine surface films: chemical characteristics, influence on air-sea interactions and remote sensing. Springer, Berlin
- Garbe CS, Handler RA, Jähne B (eds) (2007) Transport at the air sea interface – measurements, models and parameterizations. Springer, Berlin
- Gulliver JS (2007) Introduction to chemical transport in the environment. Cambridge University Press, Cambridge



- Jähne B, Monahan EC (eds) (1995) Air-water gas transfer – selected papers from the third international symposium on air-water gas transfer. AEON Verlag & Studio Hanau. <http://d-nb.info/946682526>
- Kantha LH, Clayson CA (2000) Small scale processes in geophysical fluid flows, vol 67, International geophysics series. Academic, San Diego
- Komori S, McGillis W (eds) (2011) Gas transfer at water surfaces, selected papers from the 6th international symposium. Kyoto University Press, Kyoto
- Kraus EB, Businger JA (1994) Atmosphere-ocean interaction, vol 27, 2nd edn, Oxford monographs on geology and geophysics. Oxford University Press, New York
- Liss PS, Duce RA (eds) (2005) The sea surface and global change. Cambridge University Press, Cambridge
- Robinson IS (2010) Discovering the ocean from space – the unique applications of satellite oceanography. Springer, Heidelberg
- Sarmiento JL, Gruber N (2006) Ocean biogeochemical dynamics. Princeton University Press, Princeton
- Soloviev A, Lukas R (2006) The near-surface layer of the ocean, vol 31, Atmospheric and oceanographic sciences library. Springer, Dordrecht
- Wanninkhof R, Asher WE, Ho DT, Sweeney C, McGillis WR (2009) Advances in quantifying air-sea gas exchange and environmental forcing. *Annu Rev Mar Sci* 1:213–244
- Wilhelms SC, Gulliver JS (eds) (1991) Air-water mass transfer – selected papers from the 2nd international symposium on gas transfer at water surfaces, Minneapolis Minnesota, September 11–14, 1990. American Society of Civil Engineers, New York
- Zeebe RE, Wolf-Gladrow DA (2001) CO<sub>2</sub> in seawater: equilibrium, kinetics, isotopes, vol 65, Elsevier oceanography series. Elsevier, Amsterdam

## Atmospheric Biogeochemistry

NATALIE M. MAHOWALD

Department of Earth and Atmospheric Sciences,  
Atkinson Center for Sustainable Future, Cornell  
University, Ithaca, NY, USA

### Article Outline

Glossary  
Definition of the Subject and Its Importance  
Introduction  
Cycling in the Atmosphere  
Atmospheric Modification to Other Biogeochemical  
Components

The Atmosphere as an Integrator

Future Directions

Acknowledgments

Bibliography

### Glossary

**Aerosol** A solid or liquid suspended in the atmosphere. The definition usually does not include cloud droplets, although many aerosols have water vapor on their surfaces.

**Dry deposition** Removal process for gas and aerosol species in which the species are deposited onto the lower surface due to either turbulent fluxes (overturning air) forcing the constituent to hit and stick to the surface, or from gravitational settling of aerosols. Gravitational settling is the dominant mechanism for removal for larger aerosols.

**Lifetime** The atmospheric lifetime of a constituent describes how long the constituent will remain in the atmosphere. It is typically calculated by dividing the total amount of the constituent in the atmosphere by the total flux out of or into the atmosphere. The flux can be due to atmospheric chemical reactions, and/or exchanges between other reservoirs in the earth system (e.g., land or ocean). This lifetime is an e-folding lifetime; if one starts with an initial perturbation of the constituent, the amount of the perturbation remaining after a length of time equal to the lifetime is equal to 1/e of the original value of the perturbation.

**Wet deposition** Process by which an atmospheric constituent is removed by precipitation. This is especially important for water-soluble and aerosol species.

### Definition of the Subject and Its Importance

Biogeochemistry represents the interaction of biology, chemistry, and geology in the Earth system. For many processes, an understanding of biological uptake and emission, chemical processing, and geological sequestration is necessary to resolve the sources and sinks of a particular constituent. For example, to discover the sources and sinks of atmospheric carbon dioxide, it is important to understand how biota take up carbon dioxide and chemically convert the carbon to organic carbon, and then how this organic carbon is used either to produce energy by biota or is deposited to the land

or ocean surface and can become sequestered in geological formations. Often when biogeochemistry is referred to, one refers to the nutrient cycling of important nutrients. Atmospheric biogeochemistry, as defined in its narrowest sense, is the flux of nutrients and pollutants important for biogeochemistry through the atmosphere. In its broadest sense, atmospheric biogeochemistry is the interaction of the atmosphere and atmospheric processes with the biogeochemistry of the earth system. This broader definition is used here. The hydrologic cycle is one of the most important biogeochemistry cycles, and is a worthy topic of its own. It is discussed in several other sections of this encyclopedia (e.g., Climate Change and Global Water Sustainability, Water Resource and Quality Management for Adaptation to Climate Change, Water Availability and Quality, Sustainable).

## Introduction

Atmospheric biogeochemistry is increasingly relevant for today's society, as evidenced by the changing composition of the atmosphere. Atmospheric biogeochemistry views these changes in the composition as part of a collection of changes to the earth system occurring during the anthropocene. One very serious sustainability issue at the core of atmospheric biogeochemistry is the accumulation of carbon dioxide in the atmosphere. This is a result of human emissions of carbon dioxide both directly from fossil fuel burning, and from land use change, especially tropical deforestation [1]. Interestingly, only about  $\frac{1}{2}$  of the carbon dioxide emitted into the atmosphere remains there: The rest is taken up by the land and ocean. This important negative feedback on human emissions of carbon dioxide is a result of biogeochemical cycling of carbon in the system. As human emissions of carbon dioxide grow, it is not clear that the land and the ocean will be able to keep up, and thus more of the carbon dioxide is likely to stay in the atmosphere [2]. Some studies suggest that warming itself leads to less carbon dioxide being taken up by the land and ocean [2], although this also depends on the cycling of other nutrients such as nitrogen as well [3]. Thus understanding the biogeochemistry of carbon is required to understand how much carbon can be emitted in the future.

The atmospheric transport and subsequent deposition of trace gases and aerosols can provide an important mechanism for circulation of different constituents. Whether this deposition serves as a nutrient or a pollutant usually depends on the deposition level, as well as the ability of the biota to process these compounds. Some compounds can be nutrients at low levels, but pollutants at high levels. Thus, referring to the atmospheric transport of nutrients and pollutants as "atmospheric biogeochemistry" allows for these changes in impacts depending on amount. For transport from the land to ocean, many constituents can also be carried by rivers; in the sections below, the relative importance of atmospheric versus riverine inputs is discussed. The atmosphere tends to be able to transport species more quickly and farther distances; however the mass of the atmosphere is much less than that of water, providing a constraint on the amount of constituent the atmosphere can transport. For the cases of constituent transport from the oceans to land, or upstream on land, atmospheric transport on short timescales is often the most important, as the only other mechanism would be geological changes in orography.

The final important aspect of atmospheric biogeochemistry is that it allows us to understand the composition of the atmosphere, and the fluctuation in this composition, in an earth system context. Thus understanding atmospheric biogeochemistry represents an integrative approach to understanding the atmosphere and its relationship with the land, ocean, and cryosphere.

## Cycling in the Atmosphere

Since most nutrients cycle through the atmosphere, the atmospheric portion of their biogeochemistry is important. The most important nutrients to sustain life are carbon, nitrogen, and phosphorus, but many other nutrients are required. Here some of the most important nutrient cycles for the earth system are discussed, and cycles that are being perturbed due to human activities have to be focused upon.

### Carbon Cycle

The carbon cycle receives substantial interest today because of the anthropogenic perturbation of the carbon cycle and the documentation of global warming. The carbon cycle represents the most important of the

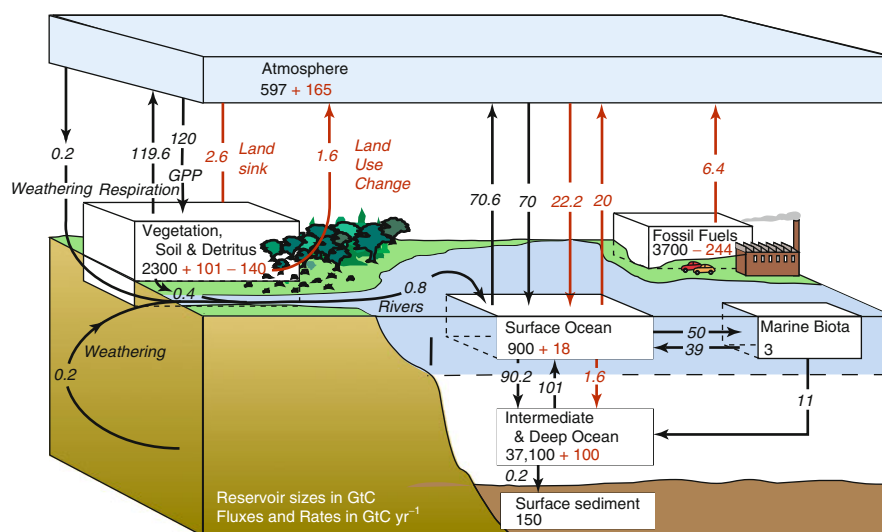
nutrient cycles, with the exception of water, for living organisms because living tissue is primarily composed of carbon. The dominant form of carbon in the atmosphere is carbon dioxide, a chemically inactive, but radiatively important species in the atmosphere. The riverine flux of carbon is 0.8 Tg/year, while the gross land and ocean fluxes with the atmosphere are closer to 120 and 70 Tg/year (the net fluxes are closer to zero). This suggests that changes in the land or ocean carbon balance will be communicated rapidly through the atmosphere.

Carbon dioxide is removed from the atmosphere, not through chemical reactions, but through huge fluxes into the land or oceans, with an atmospheric lifetime of several years (Fig. 1 from [4]). In reality, however, the land and ocean fluxes are reasonably well balanced and it is difficult for all the extra carbon to be taken up, suggesting that the lifetime of the additional carbon dioxide emitted by humans is on the order of a few hundred years [4, 5]. However, there will be some carbon (perhaps 20–35%) that stays in the atmosphere for 3–7 kyr, thus the lifetime for atmospheric carbon dioxide is ambiguous [5].

Carbon dioxide shifts between glacial and interglacial climates are one of the most robust signals from the

ice core records [6], and carbon dioxide tends to be about 80–100 ppm smaller during glacial periods than during interglacial periods (like today). These changes in carbon dioxide have not yet been explained, but are likely to be caused by a combination of colder temperatures, changes in ocean circulation, and changes in ocean productivity [7]. Indeed these changes in carbon dioxide are important forcing agents for maintaining the cold temperatures in the glacial periods [8], and thus must be an integral part of explaining the glacial/interglacial cycles over the past several hundred thousand years.

Currently, carbon dioxide emissions are increasing, and while preindustrial values of carbon dioxide were about 280 ppm, current values are over 380 ppm [4]. The reason for these changes in carbon dioxide values are due to human emissions of carbon dioxide release from combustion of fossil fuels and deforestation or other land-use change process. Carbon dioxide emissions currently (2008) are estimated to be about  $8.7 \pm 0.5$  PgC/year from fossil fuels, and another  $1.5 \pm 0.7$  PgC/year from land-use change (e.g., deforestation) [1]. Recent studies suggest that about 43% of anthropogenic carbon dioxide, from fossil fuel burning



**Atmospheric Biogeochemistry. Figure 1**

The global carbon cycle for the 1990s, showing the main annual fluxes in PgC/year: preindustrial “natural” fluxes in black and “anthropogenic” fluxes in red. Gross fluxes generally have uncertainties of more than  $\pm 20\%$  but fractional amounts have been retained to achieve overall balance when including estimates in fractions of PgC/year for riverine transport, weathering, deep ocean burial, etc. “GPP” is annual gross (terrestrial) primary production. Atmospheric carbon content and all cumulative fluxes since 1750 are as of end 1994 (Figure from [4])

or deforestation, remains in the atmosphere, while the remainder is taken up by the land or the ocean, with more than half going into the ocean [1]. In the future, it is likely that less of the carbon dioxide that is emitted will be taken up by the land and ocean, due to both a saturation of the sinks of carbon dioxide, as well as the impact of climate change on the carbon cycle [2].

This increasing carbon dioxide in the atmosphere may increase land productivity and change land biogeochemistry. Higher carbon dioxide levels may allow plants to better adapt to water stress and can enhance productivity, a process called the carbon dioxide fertilization effect [9]. It is unclear how strong this fertilization effect is on land, because of limitation by other nutrients or water [10, 11]; however, available ecosystem studies support the hypothesis that increased carbon dioxide will result in higher productivity [9].

Higher carbon dioxide levels will tend to cause the ocean to take up more carbon, since the atmosphere will be supersaturated with respect to the oceans, especially in colder regions. This additional carbon dioxide is likely to lead to lower pH values in the oceans, called ocean acidification [12], and reduce the ability of calcifying organisms to create skeletons [13, 14]. This has potentially huge repercussions on ocean biogeochemistry and biodiversity [13, 15].

Carbon also exists in the atmosphere in other forms. The next most important constituent in terms of atmospheric mass is methane. Methane is formed by the decay of reduced organic species under anaerobic conditions, occurring in wetlands, in geologic formations, or in the intestines of animals. Methane sources are estimated to be about 580 Tg/year, with anthropogenic sources (such as fossil fuel extraction or agriculture) representing 55–70% of the total emissions [4]. Methane is oxidized in the atmosphere by the OH radical with a lifetime of about 9 years, as well as smaller sinks of removal at the surface and photolysis in the stratosphere [4]. Methane is even more radiatively important than carbon dioxide on a per molecule basis, and thus its accumulation in the atmosphere is of concern for global warming. Methane concentrations have increased from ~650–1,770 ppb over the past couple hundred years, largely due to human activity [16]. Methane oxidation converts the methane into carbon dioxide.

Other important carbon-containing gases include carbon monoxide and non-methane hydrocarbons. Carbon monoxide is a product of incomplete combustion, and after carbon dioxide is the largest form of emission of carbon into the atmosphere (1,800–2,700 Tg CO/year) [17]. Carbon monoxide chemically reacts with the OH radical in the atmosphere, and has a lifetime of about 2 months [17]. The oxidation of carbon monoxide to form carbon dioxide represents an important sink of oxidants in the atmosphere [17]. Non-methane hydrocarbons refer to a large collection of carbon-containing compounds that are important for tropospheric photochemistry and the generation of ozone and oxidizing compounds. Non-methane hydrocarbons undergo oxidation via many mechanisms. The complete oxidation results in one mole of carbon dioxide for each carbon atom in the original molecule [17]. The lifetime of non-methane hydrocarbons vary considerably, but are usually less than the lifetime of methane. They are important for air quality, since they are precursors to ozone, and may themselves be toxic, but are not sufficiently abundant to have radiative impacts [17].

Carbon is also a component of atmospheric aerosols, but at much lower concentrations than the gases described above. These aerosols can, however, be important for radiation and chemistry, because they exist as a solid in a gas phase. Sources of carbonaceous aerosols include by-products from combustion processes, compounds emitted from vegetation, and in situ product from the condensation of non-methane hydrocarbons.

## Nitrogen Cycle

The dominant form of nitrogen in the atmosphere is molecular nitrogen ( $N_2$ ), which represents over 70% of the air. However, this form is chemically inert and not biologically available. Thus, the nitrogen cycle usually refers to the cycle of reactive nitrogen compounds. The dominant form of reactive nitrogen in the atmosphere is nitrous oxide, a long-lived gas (>70 years), which is destroyed predominately in the stratosphere by photolysis. The sources of nitrous oxide are in the soils, as by-products of nitrogen fixation and denitrification processes, which convert molecular nitrogen to reactive nitrogen and back. Humans have increased nitrous

oxide from 270 to 319 ppb, largely due to an increase in nitrogen fertilizer application [4]. Emissions of nitrous oxide are estimated to be 11 Tg/year. Nitrous oxide is a potent green house, and represents the third largest radiative forcing gas (after carbon dioxide and methane) [4].

Potentially even more important than nitrous oxide are the nitrogen oxides (NO and NO<sub>2</sub>), which currently have emissions of more than 140 Tg/year, while preindustrial emissions were thought to be closer to 30 Tg/year [18]. Nitrogen oxides are emitted during combustion process, with smaller contributions from soils and lightning. Nitrogen oxides have a short lifetime in the atmosphere (a few days), and are important for atmospheric photochemistry and air quality because of their role in producing ozone [17]. In addition, nitrogen oxides deposit onto the earth's surface as they are removed from the atmosphere, and thus move nitrogen from one region to another in large quantities (e.g., [19]).

Since preindustrial times, emissions and concentrations of ammonia have increased with current emissions of 52 Tg/year compared with an estimated 22 Tg/year in preindustrial times [18]. Ammonia is emitted predominately from natural soils and agriculture, with small contributions from oceans and energy use. Ammonia has a short lifetime in the atmosphere (a few days) and is removed through wet or dry deposition onto the surface.

Similar to carbon species, nitrogen is also contained in organic gas species and to a smaller extent in inorganic and organic aerosols. These species can act as reservoirs of reactive nitrogen (e.g., [17], and thus be important for atmospheric chemistry. The fraction of nitrogen species that are in aerosols is dependent on chemical equilibrium with other aerosols, especially the strong acid sulfate [20]. In the future, as sulfate emissions decrease, nitrogen aerosols may become more important in terms of the radiative forcing of the climate as the fraction of nitrogen in aerosol form increases [20].

Riverine inputs to oceans are thought to be 55–80 Tg/year [21], which is of the same magnitude as the fluxes of many nitrogen compounds discussed above, suggesting that the atmosphere plays an important role in the cycling of nitrogen between the land and ocean.

Because many terrestrial ecosystems are nitrogen limited, the impact of nitrogen deposition, especially from nitrogen oxides and ammonia, is thought to be large for land ecosystems [3, 22–25]. Nitrogen is also a limiting nutrient for ocean systems, and there have been large increases in nitrogen deposition to ocean systems [21]; however, because of the large reservoir of reactive nitrogen in the oceans, deposition to the oceans is not thought to substantially modify the ocean nitrogen budget [26].

### Phosphorus Cycle

Unlike carbon or nitrogen, most of the phosphorus in the atmosphere is in the aerosol form, with little phosphorus in the gas phase (e.g., [27]. As phosphorus is approximately 700 ppm in crustal material, and mineral aerosols (soil particles suspended in the atmosphere) represent a major source of aerosols, most of the phosphorus in the atmosphere is in mineral aerosols [28] at about 1.15 TgP/year. The next largest source is thought to be primary biogenic particles (0.16 TgP/year), but these aerosols are very poorly understood [28]. Combustion sources from fossil fuels, biomass burning, and biofuels represent a small source (0.07 TgP/year), while volcanoes and sea salts are thought to provide small amounts as well [28]. Thus, continental regions, especially deserts are the dominant source of atmospheric phosphorus [28].

Aerosols deposited to ocean have a limited residence time in the ocean mixed layer, before they sink to the ocean floor and become incorporated into the sediment. Only the fraction of phosphorus which is soluble or bioavailable can contribute to the phosphorus cycle of the ocean [28, 29]. The bioavailable fraction is often considered phosphate or soluble-reactive phosphorus [28], and the fraction of phosphorus in aerosols that is soluble varies from 7% to 100% [28, 29]. Transects of the North Atlantic suggest that desert dust aerosols average about 10% solubility [30, 31]; however a gradient in solubility is observed as the particles move away from the desert dust source area, suggesting that atmospheric processing may make the phosphorus in aerosols more soluble [30]. Studies in Mediterranean suggest that acidification, partly due to anthropogenic emissions of sulfur and nitrogen species, can make the phosphorus in desert dust more soluble [32]. Non-dust



phosphorus is thought to be much more soluble ( $\sim 50\%$ ), based on studies of European-derived aerosols in the Mediterranean [28, 33–36].

Phosphine gas ( $\text{PH}_3$ ) has been measured in limited studies, with concentrations of between 0.39 and 2.45  $\text{ng/m}^3$  in remote regions and up to 178  $\text{ng/m}^3$  in urban locations or near paddy fields [37–39]. While traditionally, the gas phase transfer of phosphorus has been considered negligible [28, 40], some of the values listed here are similar in magnitude as those found for phosphorus in aerosols [28]. This suggests that phosphine could be an important mechanism for transferring phosphorus, yet is not well understood. Small amounts of phosphine could be generated in soils, agricultural and industrial processes, and lightning, but phosphine in the presence of sunlight is converted to phosphoric acid [37, 38].

Because phosphorus limitation is thought to be widespread in tropical forests and savannahs [41, 42], atmospheric deposition of desert dust may play a role in the long-term viability of tropical soils. For example, it has been suggested that the atmospheric deposition of desert dust from North Africa is responsible for the maintenance of the Amazon forest [43], and that deposition from Asia is important source of phosphorus to the tropical forests in Hawaii on long timescales [44]. The atmospheric deposition of phosphorus is likely to be important in many land ecosystems on geological timescales [45]. Forest and grassland ecosystems can also lose phosphorus through the atmosphere, as primary biogenic particles or biomass burning particles contain a large proportion of the phosphorus in tropical forests [46, 47].

The ocean tends to be a net sink of phosphorus from the atmosphere [28], and since productivity in the ocean is thought to be phosphorus limited on long timescales [48] and in some regions [49, 50], the atmospheric deposition of phosphorus could be an important source. Riverine inputs of phosphorus to the oceans are much larger than the atmospheric fluxes described here (11  $\text{Tg/year}$ ) [51]; however this phosphorus is input to the oceans in the coastal regions, and it is unclear how much of this phosphorus is advected into the open oceans [52]. However, on short timescales, large increases (25%) in phosphorus deposition to oceans which could be due to human activity [28] is not thought to substantially impact ocean

biogeochemistry, because of the large reservoir of phosphorus already in the oceans [26].

### Sulfur Cycle

Sulfur is an important trace compound in the earth system, used by some microorganisms for energy. It is commonly found in rocks with smaller quantities measured in the atmosphere [27]. The atmospheric portion of the sulfur cycle is important because of the climate interactions of the sulfur species, especially in the sulfate aerosol form. The sulfur compound with the longest lifetime (5 years) and most common by mass is carbonyl sulfide ( $\text{OCS}$ ) [27]. The dominant sources (10  $\text{GgS/year}$ ) of carbonyl sulfide are the oceans and soils, with small contributions from combustion processes and the chemical oxidation of  $\text{CS}_2$  [27]. Carbonyl sulfide is destroyed by oxidation in the atmosphere, as well as by stratospheric photolysis and uptake by vegetation at the surface [27, 53]. Carbonyl sulfide is thought to contribute about half of the sulfur that forms the stratospheric aerosol layer, the other half coming from volcanic emissions [54].

Sulfur dioxide is a naturally occurring compound released from volcanoes and biomass burning ( $<10\%$  total) as well as oxidation of the naturally emitted dimethyl sulfide ( $\sim 20\%$ ), but the dominant source currently are human combustion activities with a total emission of about 91–125  $\text{TgS/year}$  [16, 18]. Once released into the atmosphere, sulfur dioxide is oxidized within a few days to form sulfate aerosols. Because sulfate is highly hydrophilic, sulfate aerosols take up water and make a larger contribution to aerosol-radiation interactions than their mass would suggest [16]. In addition these aerosols readily interact with cloud droplets to modify cloud optical properties (aerosol indirect effect on clouds) [16]. Sulfur dioxide from human emissions is responsible for a large portion of the anthropogenic radiation forcing of aerosols [16].

Dimethyl sulfide is a natural source of sulfur gas to the atmosphere from the ocean, with a magnitude of about 15  $\text{TgS/year}$ , which is readily oxidized to sulfur dioxide and then to sulfate aerosol (lifetime about 1.5 days). This is the dominant gas phase species of sulfur released from the ocean to the atmosphere. Because of sulfate aerosol interactions with clouds, it was hypothesized that biota could modulate their

temperature by modifying emissions of dimethyl sulfide [55]; however, recent studies suggest this mechanism is not a dominant mechanism, for example, under anthropogenic climate change [56].

Sea salts are 7.7% sulfate by weight [57], and represents the largest source of sulfur to the atmosphere. Approximately 10,000 Tg/year of sea salt aerosols [58] is emitted into the atmosphere, which translates to 770 Tg  $\text{SO}_4$ /year or 250 TgS/year. Sea salt aerosols reside in the atmosphere from hours to days, and tend to redeposit back onto oceans [58], and are important as cloud condensation nuclei [59].

Sulfur moves from the land to the ocean at a rate of approximately 130 TgS/year [27], which is similar in magnitude to many of the atmospheric fluxes, suggesting that atmospheric sulfur is an important part of the global biogeochemical cycle.

### Iron Cycle

Iron is required in small quantities by almost all organisms (e.g., [48, 60]), and is often considered a micronutrient. Because some regions of the oceans are iron limited [61], atmospheric deposition of iron is important for providing new iron to ocean biota [62]. Riverine inputs of iron are large, but are largely removed in coastal regions [63]. There is some evidence that the nitrogen-fixing organisms have larger iron requirements

Similar to the phosphorus cycle, iron resides almost entirely in atmospheric aerosols, predominately in desert dust. Combustion and other sources of iron are estimated to be less than 5% of the total iron in aerosols [64, 65]. Crustal material is on average about 3.5% iron [66], with some minerals having substantially higher or lower concentrations [67, 68]. However, estimates and observations suggest that desert dust aerosols vary in their iron content by only a factor of 2, suggesting that the high heterogeneity in the soils is mixed in the atmosphere [65, 69]. There are also small sources of atmospheric iron from volcanoes [70, 71], cosmic dust [72], and combustion [64].

Because atmospherically deposited dust only resides in the mixed layer of the ocean for a short time, many researchers consider the soluble fraction of the iron the most relevant for ocean biota [73]. However, which fraction of the iron is really bioavailable is not well understood, but is likely to be a small fraction of the total iron

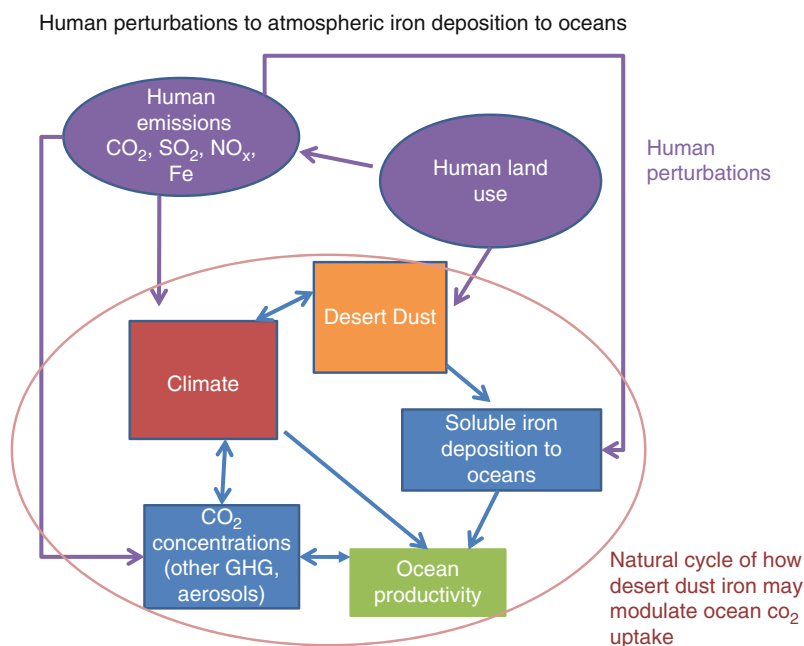
in aerosols (1–80%) (see reviews in [65, 73]). Because the soluble fraction of iron in soils is much smaller than what is observed in the atmosphere, it is thought that atmospheric processing of iron is important [73, 74]. It is likely that the acidity and insolation play a role in processing of iron [73–75]. Some iron-containing minerals are more easily solubilized than others and iron in combustions is significantly more soluble than iron in mineral aerosols; these factors complicate our understanding of iron solubility [68, 76–78] (Fig. 2 from [65]).

Some regions of the ocean are iron deficient and additions of iron can result in phytoplankton blooms [79, 80]; however, there is limited evidence showing conclusively that a case of deposition of atmospheric iron results in a bloom [81, 82]. Additional atmospheric iron deposition could also lead to increases in ocean nitrogen, because of the potential for higher iron requirements for nitrogen-fixing organisms [48, 83]. There is some evidence for this effect [84–86], although some of these effects may be due to the isotopic signature of direct anthropogenic nitrogen deposition [87]. It is likely that human changes in iron deposition are more important than the large increase in direct nitrogen deposition to the ocean, in terms of both the nitrogen budget of the ocean, as well as carbon dioxide changes [26, 88].

Estimated changes in soluble iron due to increases in desert dust sources as well as the increase in atmospheric acidification (associated with more sulfate and nitrate aerosols) could be as large as a factor of 2–4 globally averaged [65, 89]. Model results suggest this would lead to an increase in the nitrogen fixation in the oceans, and a small sink of carbon dioxide [89, 90].

### Oxygen Cycle

The atmospheric concentrations of oxygen are one of the clearest indications that life exists on this planet [27, 91]. The bulk of the atmospheric oxygen is in the gaseous state, in the form of molecular oxygen ( $\text{O}_2$ ), which represents 21% of the atmosphere. Atmospheric oxygen is changing slightly due to human activity. For every molecule of  $\text{CO}_2$  formed through combustion, one molecule of  $\text{O}_2$  is lost from the atmosphere, and while small, these changes can be detected and provide important information about the carbon cycle [92, 93].



**Atmospheric Biogeochemistry. Figure 2**

Skematic representing feedbacks between natural ocean carbon cycle, carbon dioxide concentrations, and iron inputs also shows humans could be perturbing the iron deposition (Figure from [65])

This molecular oxygen is also important as it absorbs high-energy uv light from the sun [94].

The oxygen species that exist in very small quantities in the atmosphere are extremely important for atmospheric composition, and control the oxidation of reduced species such as methane or nitrogen oxides. Ozone is produced in the stratosphere in the presence of high-energy uv light, and this ozone, as well as the process which generates the ozone, protects life from uv light at very short wavelengths [17, 94]. This also changes the thermal structure of the atmosphere, contributing to the increasing temperatures in the stratosphere [94]. Stratospheric ozone has been impacted by the release of chlorofluorocarbons, which are inert enough to reach the upper atmosphere, where they are photolyzed [95]. The largest impact has been in the polar regions, where an “ozone hole” has appeared during the hemispheric spring, although midlatitude reductions in ozone have also been observed [96]. In the troposphere, human emissions of nitrogen oxides, carbon monoxide, and non-methane hydrocarbons have led to an increase in ozone [17]. While the bulk of the ozone remains in the stratosphere, ozone is a highly oxidizing substance, so that increases in

tropospheric ozone impacts human health and materials [17]. Ozone acts as a greenhouse gas, and the observed changes in ozone, both in the troposphere and stratosphere, have impacted the radiative budget [16].

Oxidation of organic and reduced species in the troposphere is the sink for many important greenhouse gases (e.g., methane), and other important trace species in the atmosphere (e.g., sulfur dioxides, nitrogen oxides) [17]. Much of this oxidation does not occur directly by ozone, but rather by the hydroxyl radical, a by-product of ozone photochemistry in the troposphere [17]. Because this constituent reacts so quickly, it is difficult to determine its concentration directly, but it can be deduced from changes in concentrations in some of the compounds it reacts with [97]. It is unclear whether humans are increasing or decreasing the hydroxyl radical, and will continue to do so, because of the complex nonlinear chemistry [16].

### Other Cycles

Most elements are cycled in the atmosphere, and because long-range transport is much faster in the atmosphere

than on land or oceans, this transport can be important for many elements. Here the focus is on a few critical elements for global biogeochemistry. However, some elements are important for their negative impacts, especially heavy metals. For example, mercury emissions from humans impacts human and ecosystem health [98]. Some studies have suggested that atmospheric deposition of heavy metals far from source areas can adversely impact ocean biogeochemistry [99]. In addition, atmospheric deposition of acid is thought to reduce the ability of some ecosystems to retain some nutrients and to negatively impact land and aquatic ecosystems [101–103], and to enhance ocean acidification [100].

### **Anthropogenic Modifications to Atmospheric Biogeochemical Cycles**

Direct human emissions of carbon dioxide, sulfur dioxide, and nitrous oxides gases, as well carbonaceous aerosols, through the increase in combustion that humans use to generate energy, has substantially increased the sources and atmospheric concentrations of these compounds [16] (Fig. 3). In addition, changes in land use have increased carbon dioxide, methane, nitrous oxide, and ammonia emissions [16] (Fig. 3). These modifications are likely to continue into the future, with some decrease in emissions of combustion-sourced compounds estimated (Fig. 3) [101–107]. A combination of climate change and land-use change seems to have resulted in an increase in mineral aerosol over the past 100 years [89], which increases the flux of phosphorus and iron. In addition, bioavailable iron and phosphorus are likely to be even more susceptible to acidity and combustion sources; thus these compounds have changed more than simply from mineral aerosol changes over the past 100 years (Fig. 3). All of these factors have accelerated the atmospheric biogeochemical cycling of these compounds.

### **Atmospheric Modification to Other Biogeochemical Components**

#### **Atmospheric Biota**

Atmospheric aerosols are likely to be composed in part by primary biogenic particles, or particles derived from the biota, for example, plant or insect bits, pollen, bacteria, or viruses [108]. Recent estimates suggest

less than 5% of the aerosols are primary biogenic particles [109]. Some small fraction of these biogenic aerosols may contain viable living cells. While estimates of culturable bacteria are dependent on the methods used, estimates suggest that some bacteria do remain viable in the atmosphere, and can perhaps grow again if deposited onto the ground [110, 111]. Clearly, plants depend on pollen, and fungi on spores, to move genetic material from one place to another, and some fraction of this pollen and spores can be carried far from the source [112]. The amount of viable biota moving through the atmosphere and the importance of this process for global biogeochemical cycling is not well understood, but deserves future study.

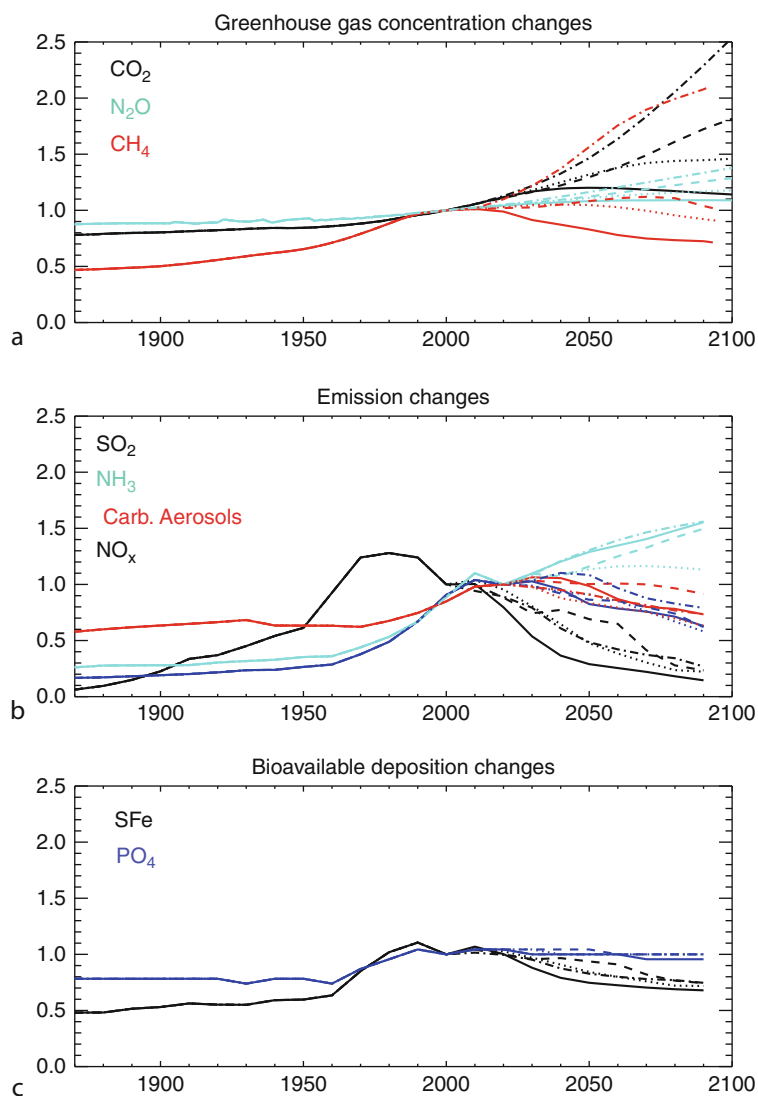
### **Climate Modification of Biogeochemical Cycles**

Changes in atmospheric biogeochemistry can modify climate, which in turn can also modify land and ocean biogeochemistry. This feedback has been used as a mechanism for understanding how carbon dioxide can be modulated on long timescales. For example, it is hypothesized that there are higher temperatures and more precipitation on land during higher carbon dioxide levels, thus increasing weathering [113]. This will reduce the atmospheric concentrations of carbon dioxide levels, acting as a feedback to prevent a runaway green house effect [113].

In the anthropocene, modeling studies have suggested that changes in anthropogenic aerosols and mineral aerosol particles can modify precipitation, incoming solar radiation, and temperature, and therefore modify land and ocean biogeochemistry enough to be seen at the global level [89, 90, 114, 115].

### **The Atmosphere as an Integrator**

The atmosphere is the one part of the earth system that is transparent in many wavelengths of light, and thus can be readily observed using remote sensing methods. In addition, we live at the interface of the atmosphere and land surface, and thus have easy access to the lower layers of the atmosphere. This means that we have much better observations of the state of many biogeochemical cycles in the atmosphere, than in the land or ocean [16]. Thus understanding atmospheric biogeochemistry provides important information for understanding global biogeochemical cycles.



### Atmospheric Biogeochemistry. Figure 3

Relative changes in green house gas concentrations (a), for carbon dioxide (black), nitrous oxide (cyan), and methane (red), relative to concentrations in 2000, based on atmospheric and ice core measurements for the past time period [127–130]. For the future, estimates are based on the representative concentration pathways estimates for four different scenarios [101–107], represented by the four lines into the future. Emission changes for important species, all normalized to 2000 values (b) for sulfur dioxide (black), ammonia (cyan), carbonaceous aerosols (red), and nitrogen oxides (blue) based on historical estimates [18], and for future estimates, the representative concentration pathways [101–107]. Deposition changes for bioavailable iron (black) and phosphorus (blue) are shown relatively to 2000 estimates (c), based on historical estimates and future estimates, using estimates of desert dust changes from [89], assuming in the future the model mean for no carbon dioxide fertilization changes in source area from [131], with iron solubility estimates from [64, 65], scaled by the sulfur dioxide emissions changes shown in (b), and for combustion soluble iron sources from [64] scaled by the black carbon emission changes included in (b). For soluble phosphorus (phosphates), desert dust estimate changes are the same as described for iron, and combustion sources of phosphorus come from [28], and are scaled by black carbon emissions, included in (b).

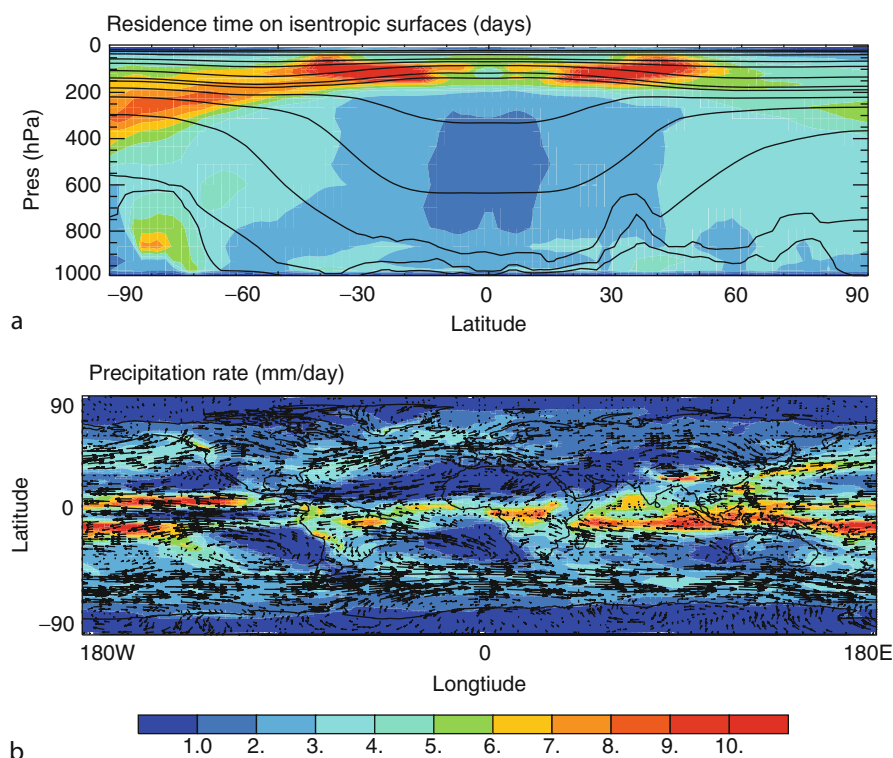


## Transport in Atmosphere

The distance a chemical travels from its source region is highly dependent on its atmospheric lifetime. Many of the biogeochemically relevant gases discussed above are long-lived gases, which will reside in the atmosphere for several months to hundreds of years (e.g.,  $\text{CO}_2$ ,  $\text{CH}_4$ ,  $\text{N}_2\text{O}$ ). Other, reactive species (such as the hydroxyl radical OH), can have very short lifetimes ( $<1$  day), and thus will not travel far from the sources. Aerosol species tend to have lifetimes between 1 day and a few weeks, and this lifetime is dependent on the size of the particle as well as whether precipitation has occurred, as this will remove particles.

For chemical species that have a relatively long lifetime ( $>$  a few days), long-range transport can occur. Once emitted into the atmosphere, constituents

mix quickly into the atmospheric boundary layer (the lowest 1–3 km of the atmosphere, where mixing is strongest), and can be transported long distances. Air parcels stay on a constant buoyancy surface (called isentropes), unless they are heated or cooled [116]. These surfaces of constant buoyancy tilt upward between the tropics and the high latitudes, because the high latitudes are colder (Fig. 4a from [109]). An air parcel emitted in the tropical boundary layer can reach the stratosphere ( $>10$  km high) in the high latitudes without being heated by moving along an isentropic surface. On average, most of the atmosphere is radiatively cooling slightly during non-stormy conditions, so, over time, parcels of air will become heavier and move downward. If the parcel encounters a storm system with clouds and precipitation, the parcel can



**Atmospheric Biogeochemistry. Figure 4**

Transport pathways of atmospheric constituents. Vertical distribution with latitude of the residence time of aerosols along a neutrally buoyant layer (isentropes), considering only vertical motion by heating in colors (a), with black lines marking surfaces of constant buoyancy (isentropes). These results are based on the 75% heating rates (warming or cooling) of the six hourly instantaneous results. Annually averaged precipitation rate (mm/day) in color, with annual mean wind vectors for 850 hPa (just above the boundary layer) (b) (Figure from [109])

experience large heating or cooling and move vertically very quickly, perhaps traveling from the boundary layer up to 10 km in less than 1 h [117]. At the same time, the precipitation within the storm system can cause wet deposition of soluble gases or aerosols. Based on atmospheric heating rates, the residence time for a parcel on a given isentropic surface varies between <1 day for the tropics to 3–6 days in midlatitudes (based on 75 percentile heating rates) (Fig. 4a). If a parcel moves away from the pole, the tilting of the neutral buoyant surfaces will direct the parcel downward vertically and vice versa.

Geographically, there tend to be particular regions in midlatitudes that have more storms (Fig. 4b), called the storm tracks. Winds in the midlatitudes (30–60°) tend to be from the west, while winds in the tropics and highlatitudes (>60°) are less uniform. Tropical aerosols are more likely to encounter storms, resulting in vertical mixing or quick removal (Fig. 4b). Transport across latitudes tends to occur less often than transport along a latitude [118]. Thus, in the mid and high latitudes, one can think of a parcel containing a constituent moving along isentropes, following the local winds, slowly cooling, and moving to a lower isentropic surface until it encounters a storm system, at which time it can be quickly removed or vertically transported.

### Inverse Modeling

Inverse modeling is the practice of using atmospheric concentrations to deduce the sources and sinks of various important compounds [119]. Because many biogeochemically relevant species have natural sources and sinks, they have sources of uncertain magnitude. Inverse modeling has become quite common for these constituents. The initial studies used simple models to best match available data [120, 121]. With time, studies focused on regional budgets but used higher temporal and spatially resolved datasets and models to resolve regional and time-varying sources as well as sophisticated statistical techniques [122–124]. These studies (called top-down studies, because they use atmospheric concentrations to constrain sources and sinks) provide important information about the sources and can provide alternative views from upscaling flux measurements or inventories (called bottom-up approaches) [125]. Current approaches often estimate parameters based on available data, instead of resolving budgets, allowing models to be tuned to best

match observations at a site, before applying that model to other regions to extrapolate to regional and global budgets [126]. Thus understanding atmospheric biogeochemistry of important compounds can facilitate a better understanding of the land and ocean biogeochemistry, using inverse modeling approaches.

### Future Directions

The study of atmospheric biogeochemistry as a field is just beginning, although some of the most important elements, such as the carbon cycle, have been studied for many years. Our knowledge, however, of even the carbon cycle is insufficient to understand how global carbon dioxide levels will respond in the future, even assuming that human emissions can be predicted [2]. Important nutrient cycles highlighted here include the carbon, nitrogen, phosphorus, sulfur, iron, and oxygen cycles. All of these cycles in some way are heavily impacted by human activities, and in addition, feedback onto our climate and environment in important ways. Many of the important innovations over the last few years have involved understanding how the different cycles interact. Thus atmospheric biogeochemistry remains an important field for continued research and innovations.

The atmosphere is unique with respect to the other components of the earth system in our ability to easily observe many layers using remote sensing. This means that the atmosphere provides a location where biogeochemistry as a whole can be better understood. Inverse methods and modeling, in addition to the many observations, may allow us to improve our understanding of not just atmospheric biogeochemical cycles, but land and ocean biogeochemical cycles.

### Acknowledgments

The author would like to thank NSF (0932946, 0832782, 0758369) and NASA (NNG06G127G), as well as Rachel Scanza for assistance on the manuscript.

### Bibliography

#### Primary Literature

1. LeQuere C, Raupach M, Canadell J, Marland G, Bopp L et al (2009) Trends in the sources and sinks of carbon dioxide. *Nat Geosci* 2:831–836. doi:10.1038/ngeo689

2. Friedlingstein P, Cox P, Betts R, Bopp L, Von Bloh W et al (2006) Climate-carbon cycle feedback analysis, results from the C4MIP Model intercomparison. *J Clim* 19:3337–3353
3. Thornton P, Doney S, Lindsay K, Moore JK, Mahowald N et al (2009) Carbon-nitrogen interactions regular climate-carbon cycle feedbacks: results from an atmosphere–ocean general circulation model. *Biogeosciences Discussion* 6:3303–3354
4. Denman KL, Brasseur G, Chidthaisong A, Ciais P, Cox PM et al (2007) Couplings between changes in the climate system and biogeochemistry. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) *Climate change 2007: the physical science bases. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, UK
5. Archer D, Eby M, Brovkin V, Ridgwell A, Cao L et al (2009) Atmospheric lifetime of fossil fuel carbon dioxide. *Annu Rev Earth Pl Sc* 37:117–134
6. Petit JR, Jouzel J, Raynaud D et al (1999) Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399:429–436
7. Sigman DM, Boyle EA (2000) Glacial/interglacial variations in atmospheric carbon dioxide. *Nature* 407:859–869
8. Hewitt C, Mitchell J (1997) Radiative forcing and response of a GCM to ice age boundary conditions: cloud feedback and climate sensitivity. *Clim Dynam* 13:821–834
9. Long SP, Ainsworth E, Rogers A, Ort D (2004) Rising atmospheric carbon dioxide: plants FACE the future. *Annu Rev Plant Biol* 55:591–628
10. Caspersen J, Pacala S, Jenkins J, Hurtt G, Moorcroft P, Birdsey R (2000) Contributions of land-use history to carbon accumulation in U.S. forests. *Science* 290:1148–1151
11. Joos F, Prentice IC, House JI (2002) Growth enhancement due to global atmospheric change as predicted by terrestrial ecosystem models: consistent with US forest inventory data. *Global Change Biol* 8:299–303
12. Caldeira K, Wickett M (2003) Anthropogenic carbon and ocean pH. *Nature* 425:365
13. Kleypas J, Buddemeier R, Archer D, Gattuso J-P, Langdon C, Opdyke B (1999) Geochemical consequences of increased atmospheric carbon dioxide on coral reefs. *Science* 284:118–120
14. Riebesell U, Zondervan I, Rost B, Tortell P, Zeebe R, Morel F (2000) Reduced calcification of marine plankton in response to increased atmospheric CO<sub>2</sub>. *Nature* 407:364–367
15. Doney S, Fabry V, Feely R, Kleypas J (2009) Ocean acidification: the other CO<sub>2</sub> problem. *Annu Rev Mar Sci* 1:169–192
16. Forster P, Ramaswamy V, Artaxo P, Bernsten T, Betts R et al (2007) Changes in atmospheric constituents and in radiative forcing. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) *Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, UK/New York, pp 130–234
17. Jacob D (1999) *Atmospheric chemistry*. Princeton University Press, Princeton, p 266
18. Lamarque J-F, Bond T, Eyring V, Granier C, Heil A et al (2010) Historical (1850–200) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. *Atmos Chem Phys* 10:7017–7039
19. Galloway J, Townsend A, Erismann J, Bekunda M, Cai Z et al (2008) Transformation of the nitrogen cycle: recent trends, questions and potential solutions. *Science* 320:889–892
20. Adams P, Seinfeld J, Koch D (1999) Global concentrations of tropospheric sulfate, nitrate and ammonium aerosol simulated in a general circulation model. *J Geophys Res* 104:13791–13823
21. Duce RA et al (2008) Impacts of atmospheric anthropogenic nitrogen on the open ocean. *Science* 320. doi:10.1126/science.1150369
22. Holland E et al (1997) Variations in the predicted spatial distribution of atmospheric nitrogen deposition and their impact on carbon uptake by terrestrial ecosystems. *J Geophys Res* 102:15849–15866
23. Townsend A, Braswell BH, Holland E, Penner J (1996) Spatial and temporal patterns in potential carbon storage due to deposition of fossil fuel derived nitrogen. *Ecol Appl* 6. doi:10.2307/2269486
24. Sokolov AP, Kicklighter DW, Melillo JM, Felzer BS, Schlosser CA, Cronin T (2008) Consequences of considering carbon-nitrogen interactions on the feedbacks between climate and the terrestrial carbon cycle. *J Clim* 21:3776–3796
25. Zaehle S, Friedlingstein P, Friend AD (2010) Terrestrial nitrogen feedbacks may accelerate future climate change. *Geophys Res Lett* 37. doi:10.1029/2009GL041345
26. Krishnamurthy A, Moore JK, Mahowald N, Luo C, Zender CS (2010) The impacts of atmospheric nutrient inputs on marine biogeochemistry. *J Geophys Res* 115. doi:10.1029/2009JG001115
27. Schlesinger WH (1997) *Biogeochemistry: an analysis of global change*. Academic, San Diego, 588 pp
28. Mahowald N, Jickells TD, Baker AR, Artaxo P, Benitez-Nelson CR, Bergametti G, Bond TC, Chen Y, Cohen DD, Herut B, Kubilay N, Losno R, Luo C, Maenhaut W, McGee KA, Okin GS, Siefert RL, Tsukuda S (2008) The global distribution of atmospheric phosphorus deposition and anthropogenic impacts. *Global Biogeochem Cycle* 22. doi:10.1029/2008GB003240
29. Graham WF, Duce RA (1982) The atmospheric transport of phosphorus to the Western North Atlantic. *Atmos Environ* 16:1089–1097
30. Baker A, French M, Linge K (2006) Trends in aerosol nutrient solubility along a west–east transect of the Saharan dust plume. *Geophys Res Lett* 33. doi:10.1029/2005GL024764
31. Baker A, Jickells T, Biswas K, Weston K, French M (2006) Nutrients in atmospheric aerosol particles along the Atlantic Meridional Transect. *Deep-Sea Res II* 53:1706–1719
32. Nenes A, Krom M, Mihalopoulos N, VanCapellen P, Shin Z et al (2010) Atmospheric acidification of mineral aerosols: a source of bioavailable phosphorus to the oceans. *Atmos Chem Phys Discuss* 11:6163–6185

33. Bergametti G, Remoudaki E, Losno R, Stiner E, Chatenet B, Buat-Menard P (1992) Source, transport and deposition of atmospheric phosphorus over the Northwestern Mediterranean. *J Atmos Chem* 14:501–513
34. Herut B, Krom M, Pan G, Mortimer R (1999) Atmospheric input of nitrogen and phosphorus to the Southeast Mediterranean: sources, fluxes and possible impact. *Limnol Oceanogr* 44:1683–1692
35. Herut B, Collier R, Krom M (2002) The role of dust in supplying nitrogen and phosphorus to the Southeast Mediterranean. *Limnol Oceanogr* 47:870–878
36. Ridame C, Guieu C (2002) Saharan input of phosphate to the oligotrophic water of the open western Mediterranean Sea. *Limnol Oceanogr* 47:856–869
37. Glindemann D, Edwards M, Kusch P (2003) Phosphine gas in the upper troposphere. *Atmos Environ* 37:2429–2433
38. Glindemann D, Edwards M, J-a L, Kusch P (2005) Phosphine in soils, sludges, biogases and atmospheric implications—a review. *Ecol Eng* 24:457–463
39. Zhu R, Glindemann D, Kong D, Sun L, Geng J, Wang X (2007) Phosphine in the marine atmosphere along a hemisphere course from China to Antarctica. *Atmos Env* 41:1567–1573
40. Graham WF, Duce RA (1979) Atmospheric pathways of the phosphorus cycle. *Geochimica et Cosmochimica Acta* 43:1195–1208
41. Vitousek P (1984) Litterfall, nutrient cycling and nutrient limitations in tropical forests. *Ecology* 65:285–298
42. Okin GS, Mladenov N, Wang L, Cassel D, Caylor KK et al (2008) Spatial pattern of soil nutrients in two southern African savannas. *J Geophys Res* 111. doi:10.1029/2007JG000584
43. Swap R, Garstang M, Greco S, Talbot R, Kallberg P (1992) Saharan dust in the Amazon Basin. *Tellus* 44B:133–149
44. Chadwick OA, Derry LA, Vitousek PM, Huebert BJ, Hedin LO (1999) Changing sources of nutrients during four million years of ecosystem development. *Nature* 397:491–496
45. Okin G, Mahowald N, Chadwick O, Artaxo P (2004) The impact of desert dust on the biogeochemistry of phosphorus in terrestrial ecosystems. *Global Biogeochem Cycle* 18:GB2005. doi:10.1029/2003GB002145
46. Graham B, Guyon P, Maenhaut W, Taylor PE, Ebert M et al (2003) Composition and diurnal variability of the natural Amazonian aerosol. *J Geophys Res* 108:4765. doi:10.1029/2003JD004049
47. Mahowald N, Artaxo P, Baker A, Jickells T, Okin G et al (2005) Impact of biomass burning emissions and land use change on Amazonian atmospheric cycling and deposition of phosphorus. *Global Biogeochem Cycle* 19:GB4030. doi:10.1029/2005GB002541
48. Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science* 281:200–206
49. Wu J, Sunda W, Boyle E, Karl D (2000) Phosphate depletion in the western North Atlantic Ocean. *Science* 289:759–762
50. Krishnamurthy A, Moore K, Zender C, Luo C (2007) Effects of atmospheric inorganic nitrogen deposition on ocean biogeochemistry. *J Geophys Res* 112. doi:10.1029/2006JG000334
51. Seitzinger S, Harrison J, Dumont E, Beusen A, Bouwman A (2005) Sources and delivery of carbon, nitrogen and phosphorus to the coastal zone: an overview of the Global Nutrient Export from Watersheds (NEWS) models and their application. *Global Biogeochem Cycle* 19. doi:10.1029/2005GB002606
52. Benitez-Nelson C (2000) The biogeochemical cycling of phosphorus in marine systems. *Earth-Sci Rev* 51:109–135
53. Crutzen PJ (1976) The possible importance of CSO for the sulfate layer of the stratosphere. *Geophys Res Lett* 3:73–76
54. Deshler T (2007) Stratospheric aerosol: measurements, importance, life cycle, anomalous aerosol. In: O'Dowd C, Wagner P (eds) *Nucleation and atmospheric aerosols*. Springer, Heidelberg, pp 613–624
55. Charlson RJ, Lovelock JE, Andreae M, Warren SG (1987) Oceanic phytoplankton, atmospheric sulfur, cloud albedo and climate. *Nature* 326:655–661
56. Kloster S, Feichter J, Maier-Reimer E, Roeckner E, Stier P et al (2007) Response of DMS in the ocean and atmosphere to global warming. *J Geophys Res* 142:G03005. doi:10.1029/2006JG000224
57. Seinfeld J, Pandis S (1998) *Atmospheric chemistry and physics*. Wiley, New York, p 1326
58. Gong SL, Barrie L, Lazare M (2002) Canadian aerosol model (CAM): a size-segregated simulation of atmospheric aerosol process for climate and air quality models 2. Global sea-salt aerosol and its budgets. *J Geophys Res* 107:4779. doi:10.1029/2001JD002004
59. Carslaw KS, Boucher O, Spracklen D, Mann G, Rae JG et al (2010) A review of natural aerosol interactions and feedbacks within the earth system. *Atmos Chem Phys* 10:1701–1737
60. Morell F, Price N (2003) The biogeochemical cycles of trace metals in the oceans. *Science* 300:944–947
61. Martin J, Gordon RM, Fitzwater SE (1991) The case for iron. *Limnol Oceanogr* 36:1793–1802
62. Fung I, Meyn SK, Tegen I, Doney S, John J, Bishop J (2000) Iron supply and demand in the upper ocean. *Global Biogeochem Cycle* 14:281–295
63. Poulton S, Raiswell R (2002) The low-temperature geochemical cycle of iron: from continental fluxes to marine sediment deposition. *Am J Sci* 302:774–805
64. Luo C, Mahowald N, Bond T, Chuang PY, Artaxo P et al (2008) Combustion iron distribution and deposition. *Global Biogeochem Cycle* 22. doi:10.1029/2007GB002964
65. Mahowald N, Engelstaedter S, Luo C, Sealy A, Artaxo P et al (2009) Atmospheric iron deposition: global distribution, variability and human perturbations. *Annu Rev Mar Sci* 1:245–278. doi:10.1146/annurev/marine.010908.163727
66. Duce RA, Liss PS, Merrill JT, Atlas EL, Buat-Menard P et al (1991) The atmospheric input of trace species to the world ocean. *Global Biogeochem Cycle* 5:193–259
67. Claquin T, Schulz M, Balkanski Y (1999) Modeling the mineralogy of atmospheric dust sources. *J Geophys Res* 104:22243–22256

68. Journet E, Desboues K, Caquineau S, Colin J-L (2008) Mineralogy as a critical factor of dust iron solubility. *Geophys Res Lett* 35. doi:10.1029/2007GL031589
69. Hand J, Mahowald N, Chen Y, Siefert R, Luo C et al (2004) Estimates of soluble iron from observations and a global mineral aerosol model: biogeochemical implications. *J Geophys Res* 109:D17205. doi:10.1029/2004JD004574
70. Frogner P, Gislason SR, Oskarsson N (2001) Fertilizing potential of volcanic ash in ocean surface water. *Geol Soc Am* 29:487–490
71. Duggen S, Croot P, Schacht U, Hoffmann L (2007) Subduction zo06Ee volcanic ash can fertilize the surface ocean and stimulate phytoplankton growth: evidence from biogeochemical experiments and satellite data. *Geophys Res Lett* 34:L01612. doi:10.1029/2006GL027522
72. Johnson K (2001) Iron supply and demand in the upper ocean: is extraterrestrial dust a significant source of bioavailable iron? *Global Biogeochem Cycle* 15:61–63
73. Jickells T, Spokes L (2001) Atmospheric iron inputs to the oceans. In: Turner DR, Huntenger K (eds) *Biogeochemistry of iron in seawater*. Wiley, Chichester, pp 85–121
74. Zhu X, Prospero J, Millero F (1997) Diel variability of soluble Fe (II) and soluble total Fe in North Africa dust in the trade winds at Barbados. *J Geophys Res* 102:21297–21305
75. Meskhidze N, Chameides W, Nenes A (2005) Dust and pollution: a recipe for enhanced ocean fertilization? *J Geophys Res* 110. doi:10.1029/2004JD005082
76. Guieu C, Bonnet S, Wagener T, Loye-Pilot M-D (2005) Biomass burning as a source of dissolved iron to the open ocean? *Geophys Res Lett* 22:L19608. doi:10.1029/2005GL022962
77. Chuang P, Duvall R, Shafer M, Schauer J (2005) The origin of water soluble particulate iron in the Asian atmospheric outflow. *Geophys Res Lett* 32. doi:10.1029/2004GL021946
78. Sedwick P, Sholkovitz E, Church T (2007) Impact of anthropogenic combustion emissions on the fractional solubility of aerosol iron: evidence from the Sargasso Sea. *Geochem Geophys Geosy* 8. doi:10.1029/2007GC001586
79. Coale KH, Johnson KS, Fitzwater SE, Gordon RM, Tanner S et al (1996) A massive phytoplankton bloom induced by an ecosystem-scale iron fertilization experiment in the equatorial Pacific Ocean. *Nature* 383:495–501
80. Boyd PW, Law CS (2001) The southern ocean iron release experiment (SOIREE) introduction and summary. *Deep-Sea Res II* 48:2425–2438
81. Bishop J, Davis R, Sherman J (2002) Robotic observations of dust storm enhancement of carbon biomass in the north Pacific. *Science* 298:817–821
82. Cassar N, Bender M, Barnett B, Fan S, Moxim WJ et al (2007) The southern ocean biological response to aeolian iron deposition. *Science* 317:1067–1070
83. Berman-Frank I, Cullen JT, Shaked Y, Sherrell RM, Falkowski P (2001) Iron availability, cellular iron quotas, and nitrogen fixation in *Trichodesmium*. *Limnol Oceanogr* 46:1249–1260
84. Michaels AF, Olson D, Sarmiento JL, Ammerman JW, Fanning K et al (1996) Inputs, losses and transformations of nitrogen and phosphorous in the pelagic North Atlantic Ocean. *Biogeochemistry* 35:181–226
85. Capone DG, Zehr JP, Paerl HW, Bergman B, Carpenter EJ (1997) *Trichodesmium*, a globally significant marine cyanobacterium. *Science* 276:1221–1229
86. Mahaffrey C, Williams R, Wolff G, Mahowald N, Anderson W (2003) Isotopic signals of nitrogen fixation over the eastern North Atlantic. *Geophys Res Lett* 40. doi:10.109/2002/GL016542
87. Knapp A, Hastings M, Sigman D, Lipschultz F, Galloway J (2010) The flux and isotopic composition of reduced and total nitrogen in Bermuda rain. *Mar Chem* 120:83–89
88. Okin G, Baker A, Tegen I, Mahowald N, Dentener F et al (2011) Impacts of atmospheric nutrient deposition on marine productivity: roles of nitrogen, phosphorus and iron. *Global Biogeochem Cycle* 25:GB2022
89. Mahowald N, Kloster S, Engelstaedter S, Moore JK, Mukhopadhyay S et al (2010) Observed 20th century desert dust variability: impact on climate and biogeochemistry. *Atmos Chem Phys* 10:10875–10893
90. Mahowald NM, Lindsay K, Rothenberg D, Doney SC, Moore JK et al (2011) Desert dust and anthropogenic aerosol interactions in the Community Climate System Model coupled-carbon-climate model. *Biogeosciences* 8:387–414. doi:10.5194/bg-8-387-2011
91. Lovelock J (1979) *Gaia: a new look at life on earth*. Oxford University Press, Oxford, UK
92. Keeling R, Shertz SR (1992) Seasonal and interannual variations in atmospheric oxygen and implications for the global carbon cycle. *Nature* 358:723–727
93. Bender M, Ho D, Hendricks M, Mika R, Battle M et al (2005) Atmospheric O<sub>2</sub>/N<sub>2</sub> changes, 1993–2002: implications for the partitioning of fossil fuel CO<sub>2</sub> sequestration. *Global Biogeochem Cycle* 19. doi:10.1029/2004GB002410
94. Hartmann D (1994) *Global physical climatology*. Academic, San Diego
95. Solomon S (1999) Stratospheric ozone depletion: a review of concepts and history. *Rev Geophys* 37:275–316
96. Chipperfield MP, Randel WJ, Bodeker GE, Dameris M, Fioletov VE, AyitÉ-LÜ DLA, Ajavon N, MÈgie G, Watson RT et al (2003) Global ozone: past and future, chapter 4. In: *Scientific assessment of ozone depletion: 2002*. WMO, Geneva, p 498
97. Prinn RG, Weiss RG, Miller BR, Huang J, Alyea FN et al (1995) Atmospheric trends and lifetime of CH<sub>3</sub>CCl<sub>3</sub> and global OH concentrations. *Science* 269:187–192
98. Selin N (2009) Global biogeochemical cycling of Mercury: a review. *Annu Rev Env Resour* 34:43–63
99. Paytan A, Mackey K, Chen Y, Lima I, Doney S et al (2009) Toxicity of atmospheric aerosols on marine phytoplankton. *Proc Natl Acad Sci USA* 106:4601–4605. doi:10.1073/pnas.0811486106



100. Doney S, Mahowald N, Lima I, Feeley R, Mackenzie F et al (2007) Impact of anthropogenic atmospheric nitrogen and sulfur deposition on ocean acidification and the inorganic carbon system. *Proc Natl Acad Sci USA* 104:14580–14585. doi:10.1073/pnas.07022181
101. vanVuuren D, den Elzen MG, Lucas P, Eickhout B, Strengers B et al (2007) Stabilizing greenhouse gas concentrations at low levels: an assessment of reduction strategies and costs. *Clim Change* 81:119–159. doi:10.1007/s10584-006-9172-9
102. Clarke L, Edmonds J, Jacoby H, Pitcher H, Reilly J, Richells R (2007) Scenarios of greenhouse gas emissions and atmospheric concentrations. Sub-report 2.1A of synthesis and assessment product 2.1 by the US climate change science program and the subcommittee on global change research. Department of Energy, Office of Biological and Environmental Research, Washington, DC
103. Smith SJ, Wigley T (2006) Multi-gas forcing stabilization with the MiniCAM. *Energy J Special Issue* #3:373–391
104. Wise M, Calvin KV, Thomson A, Clarke LE, Bond-Lamberty B et al (2009) Implications of limiting CO<sub>2</sub> concentrations for land use and energy. *Science* 324:1183–1186
105. Fujino J, Nair R, Kainuma M, Masui T, Matuoka Y (2006) Multi-gas mitigation analysis on stabilization scenarios using AIM global model. *Energy J Special Issue* 3:343–354
106. Hijoka Y, Matuoka Y, Hisimoto H, Masui M, Kainuma M (2008) Global GHG emission scenarios under GHG concentration stabilization targets. *Global Env Eng* 13:97–108
107. Riahl K, Gruebler A, Nakicenovic N (2007) Scenarios of long-term socio-economic and environmental development under climate stabilization. *Technol Forecast Soc* 74:887–935
108. Jaenicke R (2005) Abundance of cellular material and proteins in the atmosphere. *Science* 308:73
109. Mahowald N, Ward D, Kloster S, Flanner M, Heald C et al (2011) Aerosol impacts on climate and biogeochemistry. *Annu Rev Env Resour* 36. doi:10.1146/annurev-environ-042009-094507
110. Lighthart B (2006) The ecology of bacteria in the alfresco atmosphere. *FEMS Microbiol Ecol* 23:263–274
111. Burrows SM, Elbert W, Lawrence MG, Poschl U (2009) Bacteria in the global atmosphere—Part 1: review and synthesis of literature for different ecosystems. *Atmos Chem Phys* 9:9263–9280
112. Heald C, Spracklen D (2009) Atmospheric budget of primary biological aerosol particles from fungal sources. *Geophys Res Lett* 36. doi:10.1029/2009GL037493
113. Stanley S (1999) Earth system history. WH Freeman, New York, p 615
114. Jones C, Cox P, Essery R, Roberts D, Woodage M (2003) Strong carbon cycle feedbacks in a climate model with interactive CO<sub>2</sub> and sulphate aerosols. *Geophys Res Lett* 30:1479. doi:10.1029/2003gl016687
115. Mercado L, Bellouin N, Stich S, Boucher O, Huntingford C et al (2009) Impacts of changes in diffuse radiation on the global land carbon sink. *Nature* 458:1014–1018. doi:10.1038/nature07949
116. Holton JR (1989) Global transport processes in the atmosphere. In: Hutzinger O (ed) *The handbook of environmental chemistry*. Springer, New York, pp 97–144
117. Barth M, Stuart A, Skamarock W (2001) Numerical simulations of the July 10, 1996, stratospheric-tropospheric experiment: radiation, aerosols, and ozone (STERAO)-deep convection experiment storm: redistribution of soluble tracers. *J Geophys Res* 106:12381–12400
118. Bowman K, Carrie G (2002) The mean-meridional transport circulation of the troposphere in an idealized GCM. *J Atmos Sci* 59:1502–1514
119. Kasibhatla P, Heimann M, Rayner P, Mahowald N, Prinn R, Hartley D (eds) (2000) *Inverse methods in global biogeochemical cycles*. American Geophysical Union, Washington, DC, p 324
120. Bolin B, Keeling C (1963) Large-scale atmospheric mixing as deduced from the seasonal and meridional variations of carbon dioxide. *J Geophys Res* 68:3899–3920
121. Fung I, John J, Lerner J, Matthews E, Prather M et al (1991) Three-dimensional model synthesis of the global methane cycle. *J Geophys Res* 96:13033–13065
122. Gurney KR, Law RM, Denning AS, Rayner PJ, Baker D et al (2002) Towards robust regional estimates of CO<sub>2</sub> sources and sinks using atmospheric transport models. *Nature* 415:626–630
123. Bousquet P, Peylin P, Ciais P, Quere CL, Friedlingstein P, Tans PP (2000) Regional changes in carbon dioxide fluxes of land and oceans since 1980. *Science* 290:1342–1347
124. Houweling S, Kaminski T, Dentener F, Lelieveld J, Heimann M (1999) Inverse modeling of methane sources and sinks using the adjoint of a global transport model. *J Geophys Res* 104:26137–26160
125. Kasibhatla P, Arellano A, Logan J, Palmer P (2002) Top-down estimate of a large source of atmospheric carbon monoxide associated with fuel combustion in Asia. *Geophys Res Lett* 29. doi:10.1029/2002GL015581
126. Braswell B, Sacks W, Linder E, Schimel D (2005) Estimating diurnal to annual ecosystem parameters by synthesis of carbon flux model with eddy covariance net ecosystem exchange observations. *Glob Change Biol* 11. doi:1111/j.365-2486.005.00897
127. Keeling RF, Piper SC, Bollenbacher AF, Walker JS (2009) Atmospheric CO<sub>2</sub> records from sites in the SIO air sampling network. In: *Trends: a compendium of data on global change*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Dept. of Energy, Oak Ridge
128. Etheridge DM, Steele LP, Langenfelds RL, Francey RJ, Barnola J-M, Morgan VI (1998) Historical CO<sub>2</sub> records from the law dome DE08, DE08-2, and DSS ice cores. In: *Trends: a compendium of data on global change*. Oak Ridge National Laboratory: Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge

129. Etheridge DM, Steele LP, Francey RJ, Langenfelds RL (2002) Historical CH<sub>4</sub> records since about 1000 A.D. from ice core data. In: Trends: a compendium of data on global change. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge
130. Battle M, Bender M, Sowers T, Tans PP, Butler JH et al (1996) Atmospheric gas concentrations over the past century measured in air from firn at the South Pole. *Nature* 383:231–235
131. Mahowald N (2007) Anthropocene changes in desert area: sensitivity to climate model predictions. *Geophys Res Lett* 34. doi:10.1029/2007GL030472

### Books and Reviews

- Jacobson M, Charlson R, Rodhe H, Orians G (eds) (2000) *Earth system science*, vol 72. Academic, San Diego

---

## Atmospheric General Circulation Modeling

PHILIP J. RASCH

Pacific Northwest National Laboratory, Richland, WA, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Clouds and Aerosols in Climate Models  
 Climate Forcing and Response  
 Calibration (Tuning) and Evaluation of GCMs  
 Remaining Challenges in Modeling the Atmosphere  
 Acknowledgments  
 Bibliography

### Glossary

**Aerosols** The small (solid and liquid) particles that are suspended in the atmosphere. Aerosols have both natural (e.g., sea-salt, dust, and some organic compounds released by vegetation) and anthropogenic origins (e.g., the pollution released by power plants, cars, trucks, agricultural burning, etc.).

**Climate** The statistical description of characteristics of our environment over long periods, including properties like the mean and extreme values of value of fields like temperature, winds, and moisture.

**Climate sensitivity** Usually used to mean the change in globally averaged surface temperature  $\Delta T$  that would occur in a model if it were allowed to equilibrate to a forcing  $\Delta F$  associated with a doubling of CO<sub>2</sub>. It is sometimes used in a looser fashion to refer to the change in temperature resulting from a change in forcing.

**Feedback** A process in the climate system that can either amplify (“positive feedback”) or diminish (“negative feedback”) a change in climate forcing.

**Lapse rate** A term that refers to the vertical temperature decrease in the atmosphere. When that lapse rate exceeds certain thresholds, convective overturning can occur. Two threshold lapse rates are important. The dry adiabatic lapse rate identifies the rate at which an unsaturated parcel will cool if it is lifted adiabatically. If the environmental lapse rate is larger than the dry adiabatic lapse rate, a parcel lifted adiabatically will gain buoyancy and convective overturning can occur. When saturated air is lifted adiabatically, it will cool at a temperature-dependent rate as phase change occurs. At warm temperatures, the saturated adiabatic lapse rate is less than the dry adiabatic lapse rate. Since the atmosphere will produce overturning to reduce these instabilities associated with buoyant parcels, the lapse rate and water vapor amount play an important role in convection. The moist and dry adiabatic lapse rates explain much of the vertical temperature gradient to lowest order.

**Parameterization** The equations and computer code describing the representation of a particular physical process in a climate model, for example, the representation of convection.

**Radiative forcing** A change altering the energy budget of the climate system usually associated with changes in the atmospheric abundance of greenhouse gases and aerosols, or factors like solar variability and volcanic. These changes are expressed in terms of radiative forcing, which is used to compare how a range of human and natural factors drive warming or cooling influences on global climate.

**Subgrid scale** The behavior of a process at time and space scales that are smaller than the model can resolve.

**Tropopause** A permeable boundary separating two layers of the atmosphere: the stratosphere (a relatively stable region above) and the troposphere (a less stable region below where convective overturning often occurs). The tropopause varies quite smoothly in latitude. It is highest in the tropics (18 km) and decreases toward the poles (to 10 km or so).

**Weather** The short-term evolution of our environment.

## Definition of the Subject

This entry provides a brief introduction to the computer models of the atmosphere used for climate studies. The concepts of atmospheric forcing and response are developed and used to highlight the importance of clouds and aerosols to the climate system and the many uncertainties associated with their representation. Many processes that are important to the accurate representation of clouds and aerosols for climate are subgrid scale, and present both physical and computational challenges in atmospheric modeling. Other factors contributing to uncertainties in models are discussed, and some remaining challenges in atmospheric models are introduced.

## Introduction

This entry provides a brief description of models of the atmosphere used for climate studies. These models can be part of a coupled climate system model or ► [Coupled Climate and Earth System Models](#), as described by Gent elsewhere in the section *Climate Change Modeling and Methodology*, but they can also be used separately with prescribed values for surface fields, or simpler treatments for surface processes.

The atmospheric component of climate models can vary enormously in complexity. Simple atmospheric models based on energy balance arguments can be run on a laptop to provide rapid estimates of global, annual averaged properties of the atmosphere (e.g., surface temperature, e.g., [22]). The simpler types of atmospheric models (Energy Balance Models and

Models of Intermediate Complexity) are discussed in the entry ► [Climate Change Projections: Characterizing Uncertainty Using Climate Models](#) by Sanderson, and Knutti, and Edmonds et al.

Much more elaborate models typically used in Earth system models are capable of simulating the distributions (in space and time) of hundreds of atmospheric fields and processes, the interaction between those fields and processes, and their response to external forcing. In this entry, the focus is on the more complex form of atmospheric models. These models are frequently also called General Circulation Models (GCMs) or Atmospheric General Circulation Models (AGCMs). The term GCM will be used here. More detail is found in the textbooks by Washington and Parkinson [20], Jacobson [5], and McGuffie and Henderson-Sellers [9].

GCMs share a great many features with the weather prediction models described by Bacmeister in this section *Climate Change Modeling and Methodology*. Both use the “equations of motion” (simplified versions of the Navier–Stokes for fluid flow, coupled with thermodynamic and mass conservation equations) to describe the evolution of the atmosphere, and parameterizations, but the way the models are used, and the focus is different (discussed more below).

A nice history of climate science and the development of weather and climate models can be found in Weart [21]. The first incarnation of atmospheric models solved on computers can be traced back to the efforts of a small group of meteorologists and physicists initially lead by John von Neumann and later Jule Charney near Princeton, New Jersey, soon after World War II. That effort started with the solution of simplified versions of the equations of motion on the most powerful computers available at the time (less powerful than the laptops in use today).

The complexity of present-day models has increased enormously, and large communities have grown up around those models. There are perhaps a dozen comprehensive independent GCMs in use today around the world, and many more prototypes used for study and development. Those communities include scientists and computer staff engaged in the development of the basic model, including testing and evaluation, as well as scientists that use the model as a tool for investigating climate science.

An easy example of a community activity focused on a particular GCM is the Community Atmosphere Model (CAM; [http://www.cesm.ucar.edu/working\\_groups/Atmosphere/](http://www.cesm.ucar.edu/working_groups/Atmosphere/)) project, an activity started in the USA about 30 years ago. That model is part of the larger Community Earth System Modeling (CESM) Project (<http://www.cesm.ucar.edu>).

CAM is a computer code that is order 400,000 lines of FORTRAN90 code. It is capable of being run in a variety of configurations, each optimized for different purposes, for example:

- Configurations useful for paleoclimate problems. These simulations sometimes require different positions of the continents. Simulations might be performed over millennia to explore climate change in response to orbital changes, or the Sun's luminosity.
- Idealized model configurations that simplify the Earth system by assuming the underlying surface is an "aquaplanet" (a world entirely covered by water).
- Configurations able to simulate the reactive photochemistry important for understanding the evolution of ozone distributions in the middle atmosphere handling hundreds of trace constituents, and many chemical reactions between those constituents.
- Configurations useful for study the reasons for climate change arising from nature and mankind over the last two centuries, and useful for producing projections given various scenarios of change in the future.

There are many other model configurations, and this list is not inclusive.

Detailed technical notes and scientific papers describing these types of applications can be found on the web sites mentioned above. The model is "open source," and it can be downloaded and run by anyone (provided a sufficiently powerful computer is available to perform a desired calculation). Archives of previous simulations are also available for download and examination.

The CESM and CAM projects are perhaps the largest of the GCM modeling activities in existence today, and so it provides an easy example to discuss and

describe, but perhaps a dozen other activities around the world have similar capabilities.

The rest of this entry discusses the "generic" characteristics of this kind of climate model.

Global GCMs are generally run at resolutions resolving horizontal features larger than hundreds of kilometers and vertical variations of a few hundred meters. These resolutions are somewhat lower than used by weather models, where more detail is often needed. GCMs frequently include representation of processes and variables that are neglected or treated more simply in weather models (e.g., weather models often neglect details of the evolution of aerosols, or the slow evolution of greenhouse gases that affect the Earth system over longer timescales than are important for weather). GCMs may also include external forcing terms (e.g., variations in solar fluxes, or a historic database of volcanic eruption information) that are neglected by weather models.

Weather prediction models have typically been optimized to provide information about local features of the atmosphere over shorter time periods at higher resolution. Because initial conditions for weather models are constantly being "reset" to observed values, less attention has been paid to processes that affect the simulation over longer timescales (months to years). Climate models, on the other hand, focus on a description of the subtle balances and feedbacks occurring between processes and tend to describe these relationships through statistics of their long-term behavior. For many applications, climate models ignore initial conditions (weather modelers have traditionally viewed their simulations as an initial value problem; climate modelers, as a boundary value problem). A focus on statistical properties necessarily requires "multiple samples" from a distribution, with less attention on initial conditions, and more attention on the processes that control the model equilibrium or produce a transition from one climate regime to another. These points of view are changing, as discussed below.

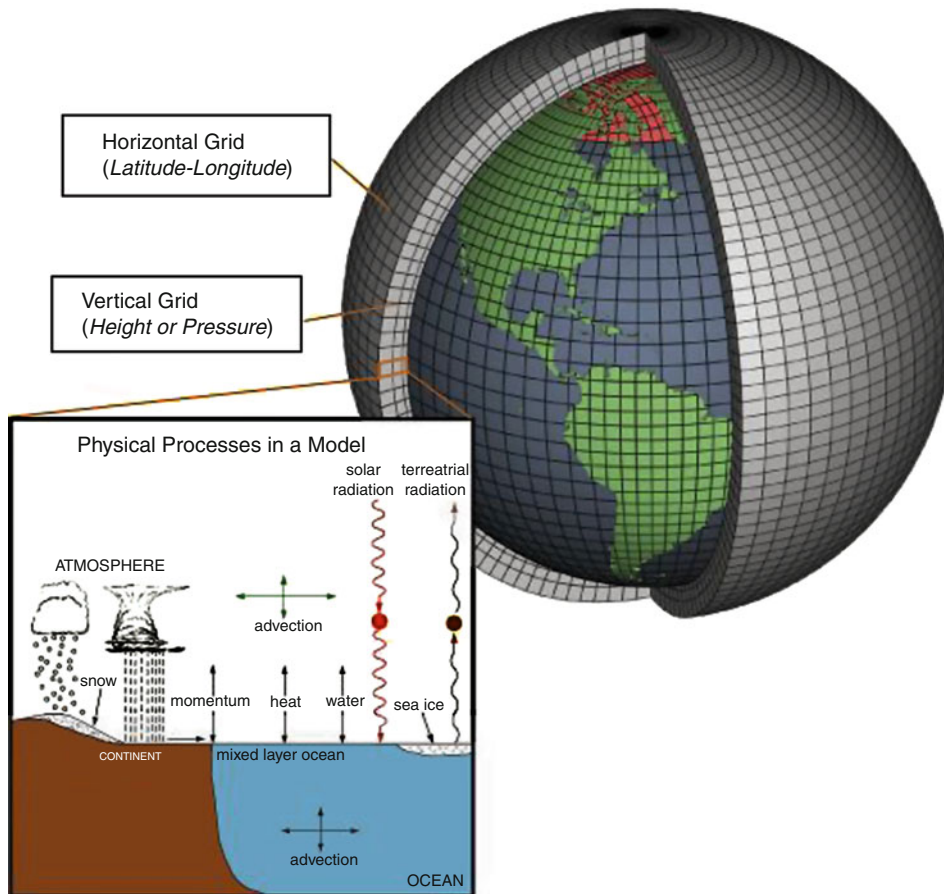
This divergence in focus between weather and climate has led to differences in model design and configuration. Climate scientists have developed models that allow simulations over centuries or millennia. Weather models provide much higher

resolution information, but those simulations are often for periods of only a few days or weeks.

GCMs being used around 2010 divide the atmosphere into columns of about a hundred kilometer on a side with 30–50 layers vertically (see Fig. 1). Most of the focus is on the atmosphere within the 40 km nearest the Earth's surface. Weather models may use columns as narrow as 10 km on a size with two to three times the number of vertical layers. Global weather models thus divide the atmosphere into volumes about 200 times smaller than climate models. Computational and accuracy constraints require that model time steps decrease in proportion to the size of the model volume. A reasonable first guess on model cost (the number of floating point operations required to complete a simulation of fixed length) is that it scales as the cube of the

model resolution. Figure 1 shows a typical type of atmospheric model grid (in this case uniform in latitude and longitude), but other discretizations are possible (see Bacmeister, this volume).

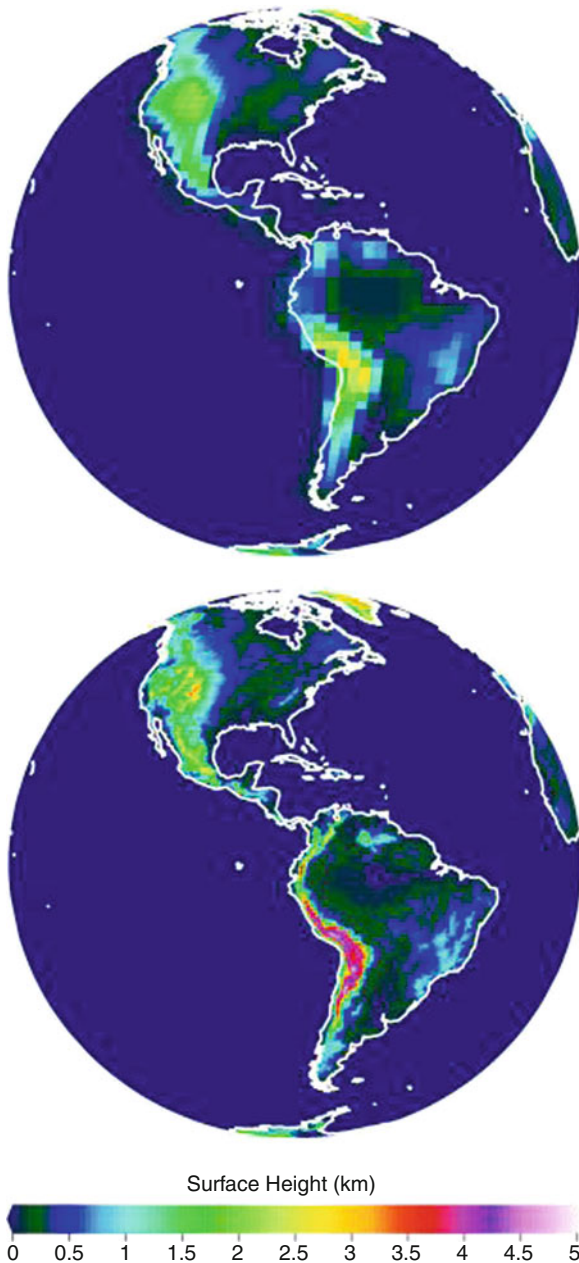
This difference in horizontal and vertical resolution produces significant differences in the way some features important to weather and climate are represented in these models. An example is shown in Fig. 2, which displays the surface topography for North and South America at a 200-km horizontal resolution typical of climate models and a 20-km resolution typical of weather models. The very sharp, small-scale topographic features like the Andes have been significantly “smoothed out” at low resolution, differing in altitude by almost a factor of 2 and spread over a much broader horizontal extent, with measurable impacts on



**Atmospheric General Circulation Modeling. Figure 1**

Typical discretization of a GCM or weather model (Figure from [http://www.oar.noaa.gov/climate/t\\_modeling.html](http://www.oar.noaa.gov/climate/t_modeling.html))



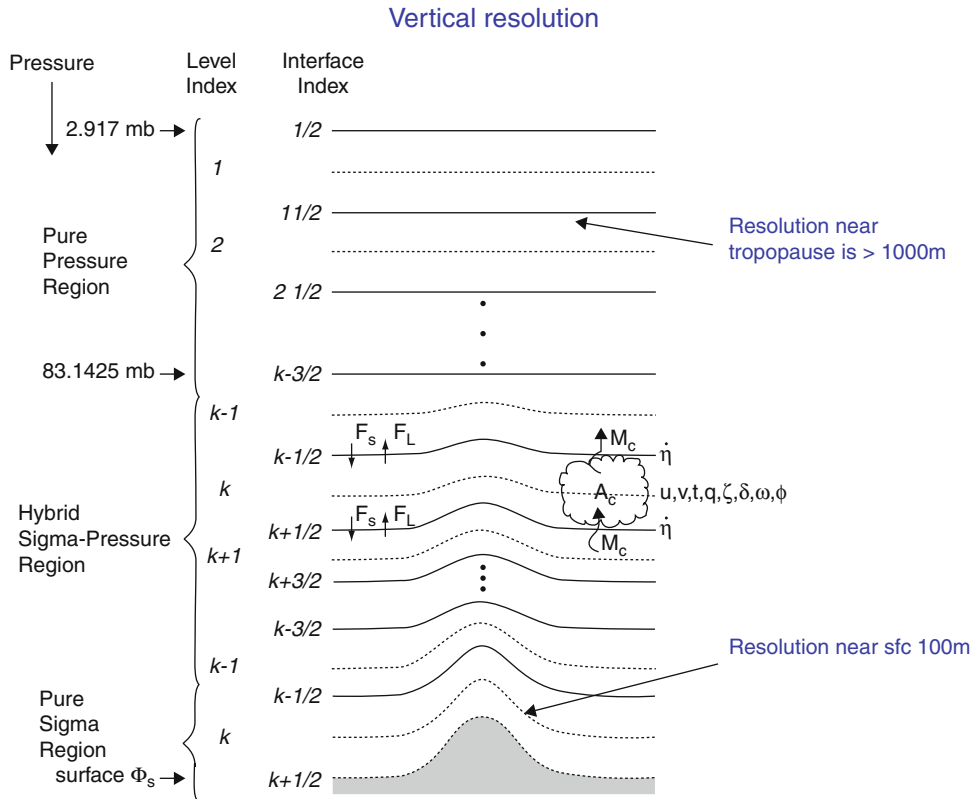


**Atmospheric General Circulation Modeling. Figure 2** The topography used by a typical “low-resolution” global atmospheric model (approximately  $2^\circ$  horizontal resolution, *upper panel*) and the high-resolution topography more typical of weather models, and next-generation climate models (about  $0.25^\circ$  resolution). Note the factor of 2 difference in height of the Andes and similar differences over the Rockies of N. America

the role of the Andes as a barrier to winds, and their role in influencing precipitation patterns. The need to resolve some features important for climate at high resolution while minimizing computational cost is one of the motivations for the development of regional climate models (see ► [Regional Climate Models](#) by Leung), or models with variable resolution meshes to put the resolution where it is needed.

Figure 3 shows a typical layer structure for an atmospheric climate model. Most models in use today rewrite their equations to employ a vertical coordinate that follows the terrain near the surface, with a gradual transition to a fixed height or pressure coordinate at higher altitudes. Model layers are generally concentrated near the Earth’s surface to deal with the complexity of processes taking place there due to boundary layer effects, terrain, interactions with surface models, and the fact that mankind lives in that region. Models typically use layers 10–100-m thick near the surface and gradually decrease that resolution to use 1–2-km-thick layers at higher altitudes. Other coordinate systems have also been considered for climate models. The equations of motion are expressed more simply with height- and pressure-based vertical coordinates, but treatment of boundary conditions is more complex. Some modeling groups have explored the use of vertical coordinates that approximately follow a material surface. These formulations result in more complex models with coordinate surfaces that can also intersect the surface of the Earth, introducing additional complexity in the treatment of boundary conditions, but the benefit is a model with much more accurate treatment of vertical transport.

Climate problems require descriptions of physical interactions at multiscale scales, leading to a very demanding challenge in physics and computational mathematics. Small-scale phenomena operating on the scale of molecules (e.g., chemistry, phase change of water, and radiative transfer) influence larger-scale features and eventually have global impacts. The physics and chemistry occurring at those small scales influence fluid motions through radiative heating and phase change to produce important phenomena like clouds with important features at scales of meters to kilometers, for example, updrafts and downdrafts. A brute force representation to treat fluid motions like up- and downdrafts would require a model



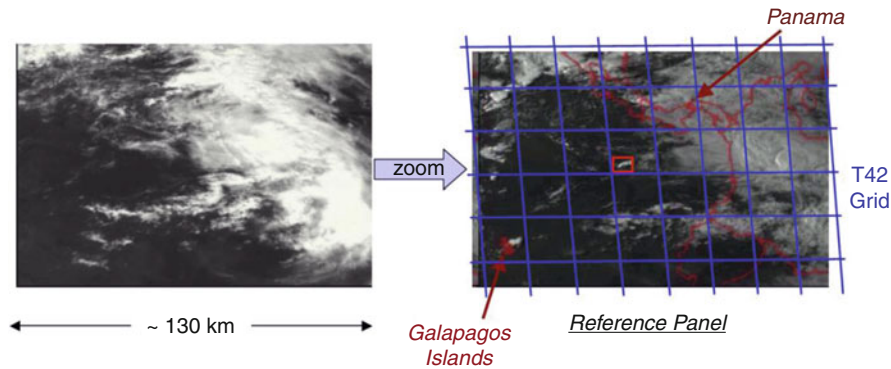
**Atmospheric General Circulation Modeling. Figure 3**

Typical distribution of layers in an atmospheric model. Top model layer will reach around 40 km for models with a focus on the troposphere and much higher for models interested in middle atmosphere problems (Figure from <http://www.cesm.ucar.edu/models/atm-cam/docs/description/description.pdf>)

that explicitly resolves those features, requiring discretizations with cells as small as a few meters on a side. It is not feasible to represent the whole globe at this resolution, and other methods are required. For this reason, many processes of climate relevance involve treatment of processes and features that are inevitably smaller than a GCM cell, or “subgrid.” Figure 4 shows a satellite image of a cloud system in the equatorial eastern Atlantic with a typical GCM grid superimposed upon it in the right panel. The cloud features are clearly below the resolution of the model. A zoomed image of a small portion of a grid cell (outlined in red) is presented in the left panel, showing important cloud features at yet finer scales.

Parameterizations have been developed in order to represent processes that are important to the atmosphere, but occur at resolutions much below the scales

the model is able to resolve (see Stensrud [16]). The equations describing the fundamental physics equations are sometimes abandoned, or simplified, through an “abstraction” to approximate that process. Sometimes those simplifications are based upon formal mathematical decompositions, like the turbulence parameterizations that depend upon “Reynolds averaging” of the equations of motion, along with appropriate choices for the constitutive equations, and “closure assumptions.” In other parameterizations, the complete equations are simplified to speed the calculation: for example, the equations of radiative heating are often approximated by assuming plane-parallel radiative transfer, and integrated over wavelength intervals to capture the essential absorption and emission for the gases and condensed species present in the atmosphere.



**Atmospheric General Circulation Modeling. Figure 4**

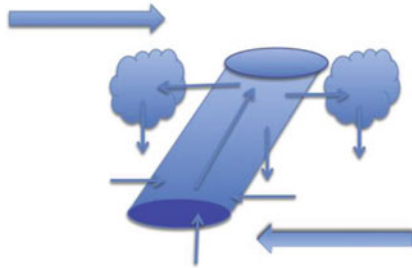
A satellite image (courtesy NASA) of the Eastern Central Pacific showing the cloud features in the context of a typical climate model resolution. The *blue lines* of the *right panel* show a superimposed grid typical of a low-resolution atmospheric model. A zoomed image of the small *red box* shown in the *right panel* appears on the left, showing even more detail

Other parameterizations are more explicitly “empirical,” employing process representations which are based upon observed behavior of the atmosphere. For example, some parameterizations for convection [2, 8] attempt to represent the overturning that occurs in the atmosphere when less dense air resides below more dense air by simply adjusting the profiles of temperature and water vapor toward prescribed profiles that agree approximately with observations. Profiles can be defined for shallow (nonprecipitating) and deep precipitating convection based on both observational evidence and theoretical considerations. So, rather than identifying a mechanism through which convection operates to reduce instabilities in the atmosphere, the parameterization makes robust statements about the “end state” of an adjustment process, and introduces empirical tendencies in the evolution equations that adjust the profile to agree with the observed profiles. This type of parameterization is more frequently used in weather prediction than in climate modeling because these empirical parameterizations may not express enough of the physical underpinning to allow inferences to be made about how, why, and where that process is important, or allow extensions to handle additional model needs. Adjustment schemes, for example, would have difficulty handling convective transport of trace species, or adjusting to changes in the fundamental physics that might be occurring as the climate

changes (e.g., the response of convective precipitation to pollution).

A third class of parameterizations resort to “process-based models.” These parameterizations replace the basic physics with a conceptual model that is assumed to mimic the processes that occur in the real world. An example of a process-based model parameterization can be seen in the use of a “bulk plume model” to represent the role of convective clouds in a model column; this type of model is used in the majority of climate models in use in 2010 (see, e.g., [23]).

In a bulk plume model, the convective overturning occurring in the atmosphere in clouds, like those seen in the right panel of Fig. 5, is envisioned to take place through an ensemble of up- and downdrafts. The updrafts are assumed to begin at the “level of free convection” (the level where a parcel lifted from the surface will be both saturated and buoyant with respect to the ambient environment). The updraft is assumed to be driven by heat released during condensation taking place in parcels within the updrafts. The condensation produced in the updraft is assumed to produce rain. The rain falling into surrounding air is assumed to partially evaporate and initiate a saturated downdraft. These up- and downdrafts carry air from one level to another, entraining air from outside the cloud in the lower part of the cloud layers to dilute the updrafts, and detraining air to the environment aloft,



**Atmospheric General Circulation Modeling. Figure 5**

The conceptual model used to produce a parameterization of a convective cloud like that seen in the figure at right. See text for details

to moisten it and redistribute heat. The ensemble of updrafts is represented by a single “bulk” updraft plume that entrains and detrains at multiple levels and a single “bulk” downdraft driven by evaporating rain to produce a conceptual model of convection like that described in the left panel of Fig. 5. The details of the representative up- and downdrafts are in turn controlled by specifications of the rate of entrainment and detrainment, how condensation, conversion of condensate to precipitation, and evaporation occurs within those up- and downdrafts, and a “closure assumption” that describes how the buoyancy generation occurring outside of the clouds is reduced by the mass fluxes within the up- and downdrafts. These parameterizations are obviously gross simplifications of the way clouds behave in the real world. The parameterizations introduce many “uncertain parameters” that require tuning to mimic the behavior of clouds in the real world.

The reader will note that the citations chosen here describe convective parameterizations written in, or prior to, the 1990s. Progress has been slow in developing better formulations for convection. Most parameterizations of convection have made progress “around the edges” by incrementally improving some aspect of the parameterization, like “closure assumptions”

(the assumptions that deter how buoyancy excesses are removed from a column) or the “microphysical formulations” (controlling the ways that condensation, conversion to precipitation, and evaporation operate within the up- and downdrafts). Cloud parameterizations are viewed by climate scientists as one of the least satisfactory components of a GCM [12]. Convective parameterizations based upon plume models have the advantage over the very simple formulations like the Betts-Miller scheme of providing a physical picture (albeit crude) of how convection works that permits the expression of conservation laws (conservation of energy, enthalpy, momentum, mass, etc.). Plume model parameterizations also allow extensions to represent interactions between aerosols and clouds, for example, or the transport of soluble and insoluble trace constituents through vigorous convection, but they still have many limitations. Recently, a new class of parameterizations has begun to be explored, in which a “nearly cloud-resolving model” is embedded within each column of a GCM (e.g., [6]). These “super-parameterizations” of clouds have their own strengths and weakness: they use equations which are very close to the original equations of motion, but those equations are solved at scales that do not really resolve cloud motions. The parameterizations also increase the cost

of the model by at least a factor of 100 over models using more traditional parameterizations so that their behavior for climate problems has not yet been thoroughly explored. Other new frameworks for cloud parameterization have also been suggested [1] that present an interesting approach to extending conventional parameterizations. There has not yet been time to evaluate the approach.

### Clouds and Aerosols in Climate Models

The accurate representation of the effect of clouds and aerosols in the atmosphere is one of the most difficult and challenging tasks in climate models for scientists at the time this entry is written.

This task is important because clouds play many roles in the atmosphere: they scatter and absorb radiant energy in both the solar (shortwave) and infrared (longwave) part of the energy spectrum, reflecting sizable amounts of sunlight (shortwave energy) back to space and thus acting to cool the planet, but they also hinder the escape of heat/energy in the longwave, and thus can warm the Earth. Clouds are also reservoirs for heat and water acting to temporarily store energy and water in condensed phases, then return it to the atmosphere at other times and locations; they are sites for important in situ atmospheric chemistry and affect photolysis rates in both clear and cloudy regions of the atmosphere by changing the actinic flux; they are regions responsible for the rapid transport of atmospheric trace constituents from the lower to the upper atmosphere through vigorous convection; and they are also entities responsible for the removal of soluble species (gases and particles) through “wet deposition” processes. And, as discussed in the previous section, they are also incredibly difficult to represent accurately and comprehensively in GCMs.

But clouds are also strongly affected by “aerosols,” the small (solid and liquid) particles with sizes less than about 10  $\mu\text{m}$  that are suspended in the atmosphere. Aerosols have both natural (e.g., sea-salt, dust, and some organic compounds released by vegetation) and anthropogenic origins (e.g., the pollution released by power plants, cars, trucks, agricultural burning, etc.). Like clouds, aerosols scatter and absorb radiant energy in both the solar and infrared part of the energy spectrum, and thus play a direct role in the energy

budget of the planet. Aerosols also affect air quality and can affect ecosystems in a number of ways (e.g., the mobilization of dust particles from deserts, their subsequent transport by winds, followed by deposition; dust deposition is believed to be a source of iron as a nutrient to ocean biota). In addition, some aerosols act as sites that facilitate the phase change of water from vapor to liquid, or ice at far lower vapor pressures than would be needed for the phase change to occur in the absence of the particles. The aerosols that act as sites for water vapor condensation to form liquid cloud drops are known as Cloud Condensation Nuclei (CCN); those that are sites for formation of ice crystals are called Ice Nuclei (IN). Different types of aerosols are more and less effective as CCN and IN, and aerosols “compete” with each other and nearby cloud drops and ice crystals for water vapor, making their interactions extremely complex and hence difficult to model (see, e.g., Seinfeld and Pandis [14] and Lohmann and Feichter [7] for complementary discussions on some of these issues).

The aerosols that become part of cloud drops or ice crystals will eventually be removed from the atmosphere when those drops or crystals get large enough to precipitate out (this is termed nucleation scavenging). Aerosols are also removed as precipitation (raindrops, snow, hail, and graupel) falls and “collects” particles along the way (termed below cloud scavenging). The treatment of aerosols thus depends intimately on the treatment of clouds in GCMs.

Aerosols and clouds thus interact in many ways. It is easy to find examples of situations where aerosols can affect the cellular structure of clouds and their reflectivity. Figure 6 shows a dramatic example of the influence of pollution from ship emissions on the brightness of low clouds near the ocean surface. Climate scientists believe that anthropogenic emissions of many aerosol types, from pollution, biomass burning, agriculture, etc., affect both the reflectivity of low clouds with impact on how clouds cool the planet, and the partial opacity of the high ice clouds that hinder the escape of heat from the planet.

The subtle interactions between clouds and aerosols, and their interactions with other components of the climate system, produce some of the largest uncertainties in interpreting the signatures of climate change over the twentieth century and complicate modelers’





**Atmospheric General Circulation Modeling. Figure 6**  
 NASA image ([http://eoimages.gsfc.nasa.gov/images/imagerecords/5000/5488/ShipTracks\\_TMO\\_2005131\\_lrg.jpg](http://eoimages.gsfc.nasa.gov/images/imagerecords/5000/5488/ShipTracks_TMO_2005131_lrg.jpg))

abilities to produce accurate projections of climate change in the future.

These issues are discussed in great detail in the fourth assessment of the Intergovernmental Panel on Climate Change (AR4, IPCC2007), and it is not possible to provide much detail here. The reader should consult AR4 and the references therein for more detail.

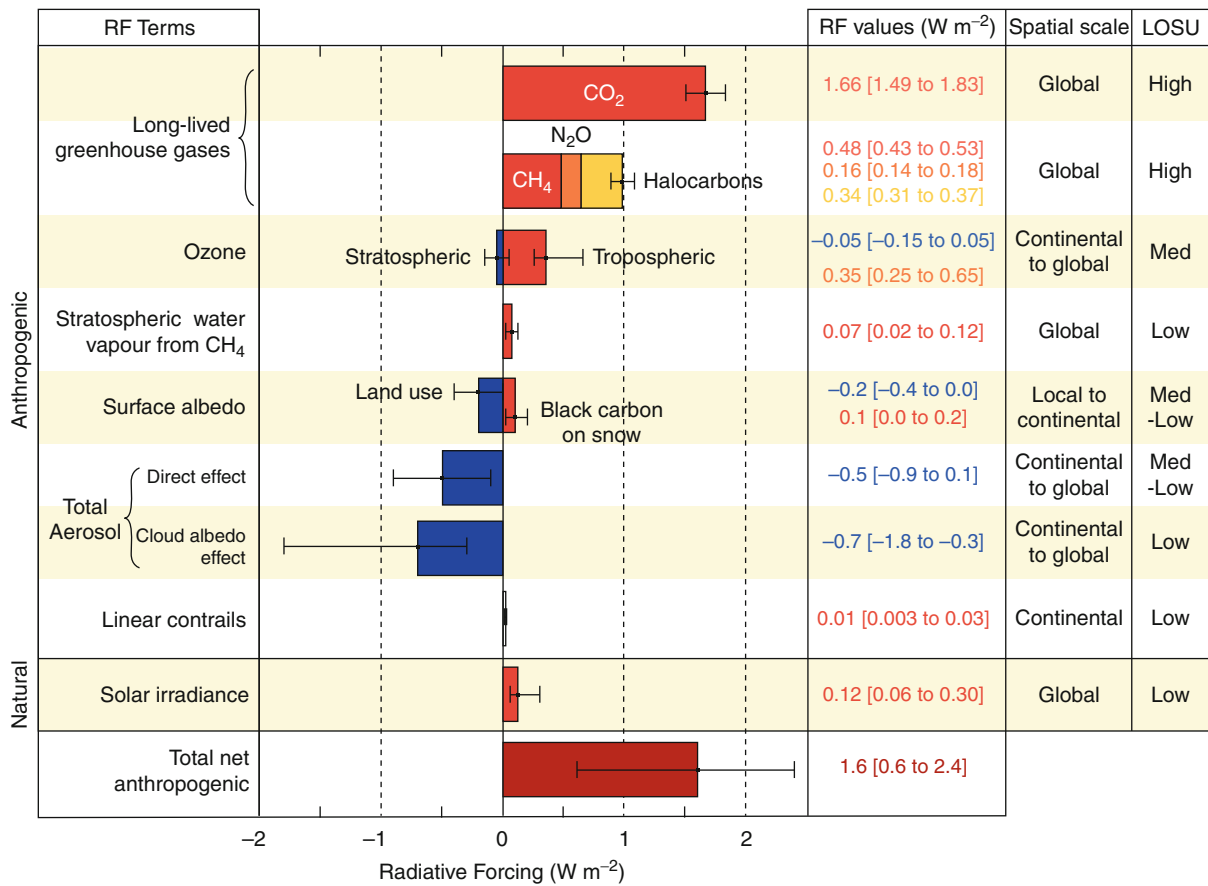
Figure 7 shows globally averaged radiative forcing estimates for various forcing agents from IPCC2007. Changes in the atmospheric abundance of greenhouse gases and aerosols, in solar radiation, and in land surface properties alter the energy balance of the climate system. These changes are expressed in terms of a “radiative forcing” ( $\text{W/m}^2$ ), a term used to compare how a range of human and natural factors drive warming or cooling trends on global climate. The three estimates related to aerosols (surface albedo, direct effect, and cloud albedo effect) are particularly noteworthy when compared to other forcing agents.

Surface albedo changes through Black Carbon deposition on snow are estimated to have a relatively small warming effect (positive forcing) on the planet. The forcing is a result of the decrease in reflectivity of the snow that occurs when the dark material is deposited on the snow surface. Since sunlight is more easily

absorbed by the darker surface in this situation, the snow melts more quickly, revealing other darker surfaces below (vegetation, dirt, sea-ice, etc.), which further increases the warming effect.

The “direct effect” of aerosols refers to the ability of aerosol to reflect or absorb sunlight as it enters the atmosphere. When the aerosols scatter sunlight back to space, the implied forcing is identified as negative (cooling) because less energy is absorbed by the planet. When aerosols absorb sunlight, they reduce the planetary albedo and warm the planet, producing a positive radiative forcing. The net radiative direct effect is estimated to be positive; models estimate that aerosols reflect more energy back to space than they absorb, but the uncertainty indicated by the horizontal whiskers is very large, and the level of scientific understanding (LOSU) is judged to be “low,” as indicated in the figure.

The aerosol “indirect effect” refers to the role that aerosols play on clouds. Increasing aerosols (e.g., from pollution) can increase the number of particles available for cloud drops or ice particles to form on by acting as CCN or IN. Those “extra” cloud drops or ice particles introduced by the additional CCN and IN compete with the ambient aerosols for water vapor, and the result is that the cloud drops and ice crystals will be smaller than they would be in the absence of pollution. Smaller drops and particles scatter sunlight more efficiently (as demonstrated by simple physical arguments and the ship tracks seen in Fig. 6), and they frequently also precipitate less efficiently, affecting cloud lifetime and areal extent. Models estimate the indirect effect to produce a very large negative forcing, but the whiskers again indicate that the uncertainty is very large, and the LOSU very low. While the processes that produce cloud brightening in the presence of additional CCN and IN are well understood, there are numerous other factors that complicate the response of cloud reflectivity enormously, and scientists know that climate models treat these other factors very crudely, and inaccurately (see Stevens and Feingold [17]). For example, increasing the number of cloud drops and decreasing their size can also cause cloud drops to evaporate more readily, so the cloud reflectivity can actually decrease, and changing the reflectivity of particular regions of a cloud system can induce changes in the cloud dynamics (the intensity and extent of up- and



**Atmospheric General Circulation Modeling. Figure 7**  
From the summary for policymakers, IPCC AR4 [4] showing the globally averaged radiative forcing estimates for various forcing agents along with uncertainty estimates and level of scientific understanding (LOSU)

downdrafts that control the precipitation and cloud areal extent; see, e.g., [19]), changing the cloud morphology and thus its radiative forcing.

The take-home message from the figure and discussion above is that aerosols of anthropogenic origin are currently estimated to have offset a substantial fraction of the positive forcing (heating) produced by increasing greenhouse gas concentrations over the twentieth century, but that result is very uncertain, and the level of understanding is low. These uncertainties in the estimates of the role of aerosols, clouds, and their interaction are strongly influenced by the deficiencies in the model representation of these processes, and by remaining deficiencies in our understanding of how these processes act.

Since it is not known how much of the twentieth century climate change (e.g., changes in surface temperature or precipitation) should be assigned to aerosol forcing and how much to the changes in greenhouse gas concentrations, it makes it much more difficult to interpret the system response to that forcing, and using understanding developed from simulations of past climate, provide accurate projections of how climate will change in the future.

**Climate Forcing and Response**

Climate change can be thought of as the response by the Earth system to the combination of externally imposed natural (e.g., solar variability and volcanic activity that changes the albedo of the planet) and

anthropogenic forcings (e.g., greenhouse gases and aerosols), modulated by the internal model processes that allow the system to adjust to the imposed forcing. The response is strongly influenced by internal feedbacks within the Earth system. Negative feedbacks will increase the rate of cooling in the presence of positive external forcing (warming). Positive feedbacks can amplify that warming (see the discussion in the entry ► [Coupled Climate and Earth System Models](#) by Gent). The relative importance of positive and negative feedbacks in the climate system controls the amplitude of climate change produced from a given amount of external forcing.

There are a variety of ways to characterize the ratio of forcing to response. One convenient measure is “climate sensitivity.” Climate sensitivity is sometimes expressed in terms of a feedback parameter  $\lambda$  (expressed in  $\text{W m}^{-2} \text{K}^{-1}$ ) or its inverse,  $1/\lambda$ . Often-times, the sensitivity is expressed as the change in globally averaged surface temperature  $\Delta T$  that would occur in a model if it were allowed to equilibrate to a forcing  $\Delta F$  associated with a doubling of  $\text{CO}_2$ :

$$\Delta T_{2\times\text{CO}_2} = \lambda^* \Delta F_{2\times\text{CO}_2}$$

It is generally assumed that  $\lambda$  can be expressed as the sum of a sequence of feedbacks  $\lambda_i$  where  $i$  indicates the process responsible for the feedback:

$$\lambda = \sum \lambda_i$$

The equilibrium change in surface temperature is a somewhat arbitrary measure of climate response, and other measurements have also been explored. The oceans have a very large heat capacity, and the rate of transport of heat into the deep oceans is very slow, which means that it would take a very long time (thousands of years) to reach an equilibrium. It is possible, with clever analytic methods, to estimate this equilibrium sensitivity, and other definitions are also used (e.g., the “transient climate sensitivity”) to describe the ratio of forcing to response. The different ways are not critical for this discussion, and the equilibrium definition is followed here.

This value is estimated in IPCC2007 (AR4) as “likely to be in the range 2–4.5°C with a best estimate of about 3°C, and is very unlikely to be less than 1.5°C. Values substantially higher than 4.5°C cannot be

excluded, but agreement of models with observations is not as good for those values.”

The first, largest, and perhaps easiest feedback to describe is the so-called Planck Feedback or Planck response, which describes the increase in emission that will occur as temperature increases due to a positive forcing. If one assumes from theory and detailed radiative calculations that the change in forcing  $F$  produced by a change in  $\text{CO}_2$  from concentration  $C_0$  to  $C$  (e.g., Myhre [25]) can be written as

$$F = k^* \ln (C/C_0) \text{ where } k \sim 5 \text{ W m}^{-2}$$

then a doubling of  $\text{CO}_2$  will result in an increase in the forcing of roughly  $4 \text{ W m}^{-2}$ . That positive forcing will tend to increase surface temperature. If one then assumes that (1) the emission temperature of the planet is proportional to the surface temperature, (2) the planet radiates energy in proportion to the Stefan-Boltzman equation ( $\sigma T^4$ ), and (3) no other atmospheric properties (clouds, water vapor, etc.) change, then model calculations indicate that  $\lambda$  is about  $-3.2 \pm 0.1 \text{ W m}^{-2} \text{K}^{-1}$  where the sign is chosen negative to indicate that the feedback is negative. And it follows that an increase of  $4 \text{ W m}^{-2}$  in forcing will result in an approximate change of 1 K in surface temperature.

This feedback estimate is robust, with very little uncertainty, and it is much lower than AGCMs report. The higher values of climate sensitivity are a result of the amplification of the response by other positive feedbacks that exist in the climate system (see, e.g., [4, 15]).

The largest positive feedback is believed to be the water vapor feedback: observations and models indicate that relative humidity (the ratio of ambient water vapor to the saturation value at a given temperature) remains approximately constant as temperature changes, particularly at high altitude (5–20 km). Therefore, an increase in temperature (e.g., from  $\text{CO}_2$ ) will produce an increase in water vapor, which is the strongest of greenhouse gases. That increase in water vapor hinders the escape of energy, and the warming is amplified. IPCC2007 estimated  $\lambda_{\text{WV}}$  to be about  $1.8 \text{ W m}^{-2} \text{K}^{-1}$ .

Another important feedback is the “lapse rate feedback”  $\lambda_{\text{LR}}$ . When estimating the Planck Feedback, it was assumed that temperature change was constant

(in latitude, longitude, and altitude). But it is known that the atmosphere will not produce a uniform change in the presence of a new forcing. The observed atmospheric lapse rate (vertical temperature gradient, see Glossary) roughly follows a “moist adiabatic lapse rate” in the tropics. It is temperature dependent and decreases more rapidly at cool temperatures than warm. So a temperature increase introduced near the surface in the tropics will produce adjustments that approximately follow a moist adiabatic lapse rate, and the perturbation will amplify with altitude. Since emission of infrared radiation varies with temperature, it will be more efficient as temperature increases, producing a negative *lapse rate feedback* that weakens the greenhouse effect. Model studies indicate that  $\lambda_{LR}$  has to be about  $-0.84 \text{ W m}^{-2} \text{ K}^{-1}$ .

It is interesting to note that models suggest that the water vapor feedback and lapse rate feedback are strongly (negatively) correlated, and the agreement by models on the sum of these two feedbacks is much more robust than the individual components:  $\lambda_{LR+WV} = 0.95 \pm 0.1 \text{ W m}^{-2} \text{ K}^{-1}$ .

The surface albedo feedback occurs because an increase in surface temperature due to a positive forcing can melt surface snow and ice. A decrease in ice and snow reduces surface reflectivity, allowing more energy to be absorbed at the surface, producing further warming, and further reducing snow and ice. Soden and Held [15] and IPCC estimate the surface albedo feedback  $\lambda_{alb}$  to be  $0.26 \pm 0.08 \text{ W m}^{-2} \text{ K}^{-1}$ .

The last feedback to be discussed is the cloud feedback. A variety of observational, theoretical, and modeling studies suggest that low clouds tend to cool the planet by reflecting sunlight back to space. High ice clouds not only reflect sunlight back to space but also have a “greenhouse effect” and hinder the escape of longwave energy to space. Observational and modeling studies indicate that the net effect of high and low clouds is to cool the Earth (cloud reflection dominates the longwave trapping of energy). But it is by no means clear how cloud radiative forcing will change in the presence of external climate forcing. Clouds are so varied and complex that fewer clear general statements emerge to guide inferences about the sign and amplitude of their feedback processes. This is an area of very active research. Soden and Held found that all the GCMs used for AR4 had a positive cloud feedback

( $0.68 \pm 0.37 \text{ W m}^{-2} \text{ K}^{-1}$ ), even though half the models exhibited a reduction in net radiative forcing in response to a warmer climate. They concluded that change in cloud forcing itself is not a reliable measure of the sign or absolute magnitude of cloud feedback due to noncloud feedbacks on the cloud forcing. Note that the uncertainty in feedback amplitude from clouds is about three times larger than that found for the lapse rate + water vapor feedback, or the albedo feedback.

The combination of positive and negative climate feedbacks produces the likely warming range of  $2\text{--}4.5^\circ\text{C}$  for a  $\text{CO}_2$  doubling cited in IPCC2007. Climate skeptics contend that the planet is unlikely to warm as much as GCMs predict by arguing that negative feedbacks are missing or underestimated or positive feedbacks overestimated in GCMs. These criticisms frequently appear in informal venues (blogs, the popular press, and elsewhere). When they are submitted to peer-reviewed refereed publications, they are taken seriously and scrutinized further. To date, the feedbacks described here have been found to be robust and dominant, and estimates of the range and distribution of climate quite robust.

### Calibration (Tuning) and Evaluation of GCMs

There are still aspects of the atmosphere that are poorly characterized, and many processes remained crudely represented in GCMs. Even in situations where the correct physics is known, it is often too expensive to include the knowledge with brute force techniques, and approximations must be employed. Both lack of understanding and process approximation lead to uncertainties, and these uncertainties produce significant variations in model formulations adopted by groups around the world.

Multiple alternative parameterizations may also exist for a particular process (e.g., convection). Even in the event that a certain configuration of parameterizations is selected, there are many “uncertain” parameters within that configuration. The choices adopted for those uncertain parameters can have strong impacts on the behavior of the model.

So, substantial resources in the modeling community are invested in evaluating the behavior of the model in the presence of these uncertainties, and in selecting the parameterizations to be used in the model,

or the values of the uncertain parameters to be used in subsequent simulations. The process of choosing the parameter values is known as “calibration” or “tuning.” Model tuning has historically been performed in a series of stages, and it is sort of an “Art” that requires a lot of insight by participating scientists, and perhaps multiple repetition of those stages.

One obvious method of tuning is to compare the behavior of a model, or a model component to observations that strongly constrain the process and adjust the parameters until the simulation agrees to some tolerance with observations. This tuning procedure can be considered an optimization problem, and it occurs frequently during model development.

For example, a parameterization of deep convection might first be tuned to make sure that it produces approximately the observed rate of precipitation, and observed tendencies of water vapor, and temperature, at a particular location and time period (the outputs) where deep convection occurs frequently, given a set of measurements of the atmospheric state (the inputs).

After this tuning, parameterization evaluation can begin by looking at its behavior at other locations and times not used for tuning. Some insight is gained during this process about reasonable values for parameter, and parameter sensitivity to variations of tunable parameters, and inputs. Sometimes “single-column model” versions of the GCM are used for this purpose. This is a version of the GCM in which a single-column model is isolated, and all lateral fluxes of information supplied from observations to identify the model behavior in strongly prescribed situations.

The model can also be tuned to optimize relationships after processes equilibrate with each other; the statistical characteristics of the model can be evaluated and compared with statistical properties of the atmosphere. For example, model top of atmosphere energy fluxes could be compared to observed energy fluxes. Small changes in tunable parameters may be performed to assure that the approximate global averages of top of atmosphere fluxes are similar to observed values without significantly degrading the agreement with observations at the process level.

The model can also be compared with observations for fields, processes, or situations that were not part of the calibration process. [Figure 8](#) provides one example

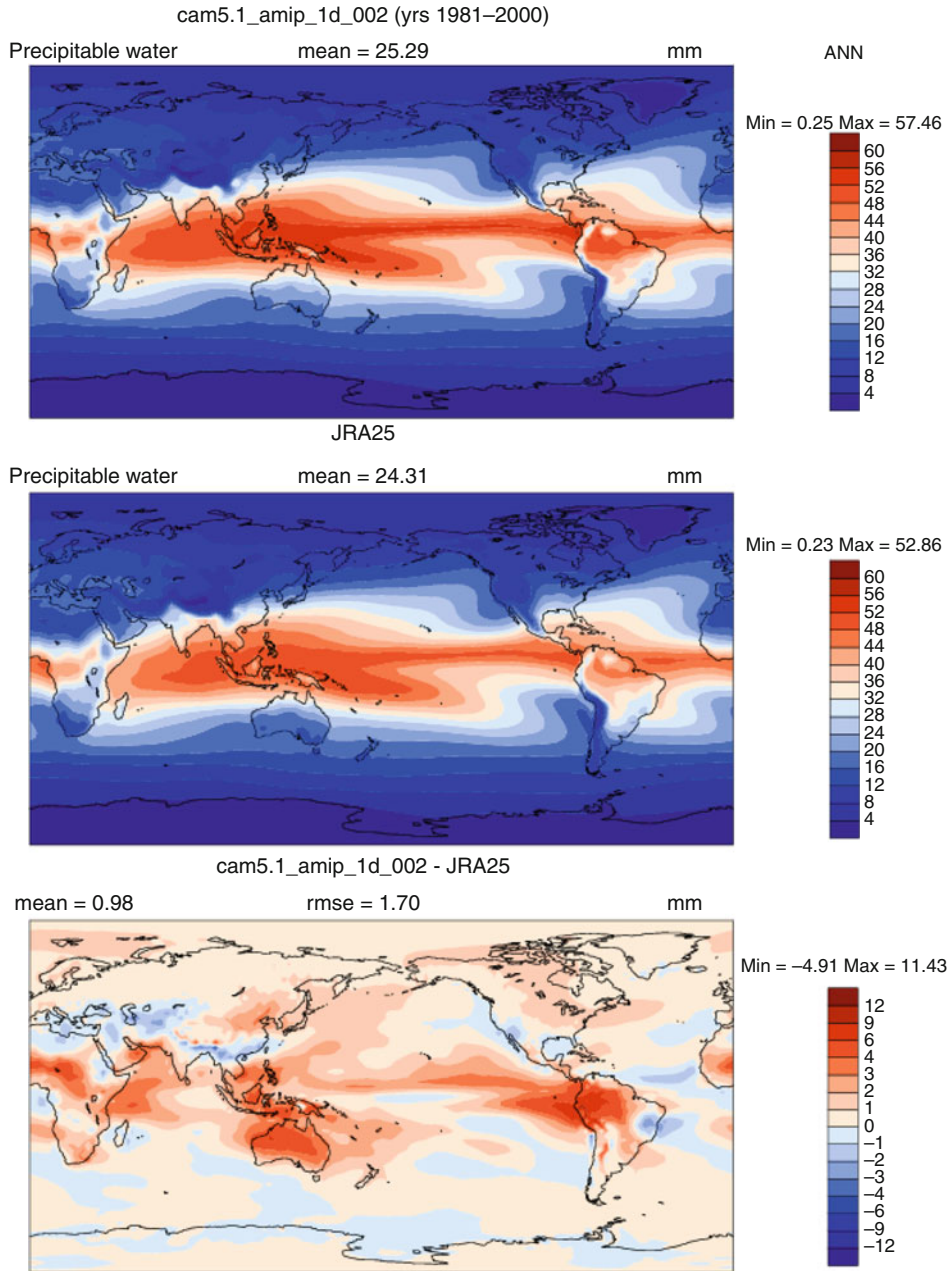
of this kind of evaluation showing differences between model annually averaged column-integrated water vapor in a long simulation of a climate model (top panel), compared to an estimate of the corresponding observed value (middle panel) and the difference. Column-integrated water is actually a pretty difficult field to observe and observational uncertainty quite high, but a variety of independent methods are available to provide estimates, and the signatures seen in the difference field are quite robust to choices of the observational dataset used. This particular model is quite moist in the tropics.

Another interesting evaluation and calibration method is the use of a climate model in “weather forecast” mode. The climate model is started from initial conditions from archives of meteorological conditions produced from a forecast center and run for a brief (few day) simulation. These kinds of simulations are often called “hindcasts.” The evolution of differences between the short simulations and observed fields provides information about how and why errors develop in the model. Individual processes can be examined to help in identifying the role of each process in driving the error growth.

The strategy can also be extended to longer periods (months or seasons) as part of a strategy sometimes called “seamless prediction” [10]. Models are viewed as representing processes that operate over different timescales. Fast processes (e.g., clouds) respond and produce measurable signatures to forcing on timescales of a day or so, systematic changes to heating rates can perturb planetary-wave structures in the atmosphere on the timescale of 10 days, and ocean and land surface react to changes in wave structure on the timescale of 100 days. On timescales of a thousand days and longer, anomalies will produce modifications to the cryosphere and biosphere. Atmospheric models (without coupling to oceans, ice, and the biosphere) can thus react and produce meaningful information on forcing and response through processes like clouds and planetary waves. Longer timescale phenomena need to be evaluated with coupled models. Seamless prediction paradigms ask that models be evaluated and improved on each timescale to provide increasing confidence in the model.

Once an “optimized” or “calibrated” model configuration has been found, it can be compared with





#### Atmospheric General Circulation Modeling. Figure 8

An example of the difference in the vertically integrated amount of water vapor simulated in a climate model run for 20 years with present-day forcing, compared with a reanalysis produce (an estimate produced from observations). The difference is shown in the *lower panel*

other models, or other configurations of the model to quantify its sensitivity to changes in forcing, to identify the reasons that it responds the way it does, and to put that sensitivity in context

compared to the behavior of other models. This process is one component of an evolving field of climate change science called “Uncertainty Quantification.”

## Remaining Challenges in Modeling the Atmosphere

There are many situations where scientists understand the fundamental physics of a physical process much more thoroughly than they can afford to express in a climate model, and scientists know that it is important to represent those processes more accurately. These are, in a sense, “known unknowns” in the parlance of Donald Rumsfeld. Scientists know that the parameterizations of clouds and aerosols used in climate models today can be formulated more accurately, and they know (in principle) how to do it. The challenge is (1) representing fields and processes that are truly important for the relevant climate problem and (2) expressing knowledge about that process in a computationally tractable way. This requires both scientific study and computational work. Many of these improvements can be achieved by clever revisions to the computational representation of critical processes in models. Scientists are rewriting model components to make more efficient use of evolving computer architectures that are providing enormous increments in computational resources, and developing more accurate and efficient computation methods to represent those processes. It may be possible in a few years to “brute force” problems that in the past required clever but potentially inaccurate approximations.

While some efforts to create a global cloud-resolving model have been made, their computation cost is at least a million times higher than models run at current climate resolution (e.g., [13]). It is currently a challenge just to complete a single annual cycle with models at this resolution. Scientists are also rewriting models to support “variable resolution” to allow models to provide higher resolution in areas that require it. One of the immediate challenges in this situation is that many of the approximations currently employed in atmospheric parameterizations are “scale dependent,” that is, the quantitative behavior of the parameterization changes behavior as the resolution changes. Current parameterizations are typically “retuned” for different resolutions, and any cloud parameterization that adapts to resolution changes is not yet known, although scale-aware parameterizations of subgrid-scale propagation of waves and wave

breaking do exist. It is desirable that parameterizations become “scale aware” in the sense that they adapt to the resolution they are used at. Parameterizations of subgrid-scale processes should recognize that the definition of the phenomena they are being asked to resolve changes as the resolution changes [1]). Scale-aware parameterizations and the model equations that they are embedded in should reduce to “fundamental physics” as model resolution approaches the resolution of the process being represented. This is an area of active research.

There are also challenges for climate models in treating “unknown unknowns,” or “missing physics.” These are phenomena that scientists believe may be important to represent, but that they do not know about or are very unsure of. An example of this situation is the role of Organic Aerosols on climate. These aerosols are difficult to measure and model. They are semivolatile; the aerosols condense, evaporate, and evolve chemically with atmospheric dilution. Aerosol scientists now believe that these aerosols play a much larger role in the climate system than was originally recognized, but they only have hints as to the complex chain of events that produce these aerosols. Precursor emissions, chemical evolution, and interaction with other aerosols and atmospheric trace species are complex, and they are treated very simply in climate models today.

Finally, there are many ways that climate science can be advanced by improvements in methodology. Ensembles of model simulations can be used to characterize model uncertainty. Model intercomparison activities like AMIP [3] and CMIP5 [18] continue to improve and allow more robust evaluation of the strength and weaknesses of climate models and their ability to represent the real world and their performance as a tool for understanding climate change.

## Acknowledgments

I would like to thank Sarah Fillmore for her editorial help and my colleagues at the Pacific Northwest National Laboratory and the National Center for Atmospheric Research for their willingness to share their expertise, knowledge, and model results with me over many years.

## Bibliography

1. Arakawa A (2011) Toward unification of the multiscale modeling of the atmosphere. *Atmos Chem Phys* 11:3731–3742. doi:10.5194/acp-11-3731-2011, [www.atmos-chem-phys.net/11/3731/2011/](http://www.atmos-chem-phys.net/11/3731/2011/)
2. Betts AK, Miller MJ (1993) The Betts–Miller scheme. In: Emanuel KA, Raymond DJ (eds) *The representation of cumulus convection in numerical models of the atmosphere*. American Meteorological Society, Boston
3. Gates WL et al (1999) An overview of the results of the atmospheric model intercomparison project (AMIP I). *Bull Atmos Sci* 80:29–55
4. IPCC (2007) Summary for policymakers. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) *Climate change 2007: the physical science basis. contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, UK/New York
5. Jacobson MZ (2005) *Fundamentals of atmospheric modeling*, 2nd edn. Cambridge University Press, New York, p 813
6. Khairoutdinov MF, Randall DA, DeMotte C (2005) Simulations of the atmospheric general circulation using a cloud-resolving model as a super-parameterization of physical processes. *J Atmos Sci* 62:2136–2154
7. Lohmann U, Feichter J (2005) Global indirect aerosol effects: a review. *Atmos Chem Phys* 5:715–737. doi:10.5194/acp/2005-5-715, [www.atmos-chem-phys.org/acp/5/715/](http://www.atmos-chem-phys.org/acp/5/715/)
8. Manabe S, Smagorinsky J, Strickler RF (1965) Simulated climatology of a general circulation model with a hydrologic cycle. *Mon Weather Rev* 93:769–798
9. McGuffie K, Henderson-Sellers A (2005) *A climate modeling primer*. Wiley, New York
10. Palmer TN, Doblas-Reyes FJ, Weisheimer A, Rodwell MJ (2008) Towards seamless prediction: calibration of climate-change projections using seasonal forecasts. *BAMS* 89: 459–470
11. Phillips TJ et al (2004) Evaluating parameterizations in general circulation models: climate simulation meets weather. *Bull Am Meteorol Soc* 85:1903–1915
12. Randall D, Khairoutdinov M, Arakawa A, Grabowski W (2003) Breaking the cloud parameterization deadlock. *Bull Am Meteorol Soc* 84:1547–1564
13. Satoh M, Matsuno T, Tomita H, Miura H, Nasuna T, Iga S (2008) Nonhydrostatic icosahedral atmospheric model (NICAM) for global cloud resolving simulations. *J Comput Phys* 227: 3486–3514
14. Seinfeld JH, Pandis SN (1997) *Atmospheric chemistry and physics*. Wiley, New York
15. Soden B, Held I (2006) An assessment of climate feedbacks in coupled ocean–atmosphere models. *J Climate* 19:3354–3360
16. Stensrud DJ (2007) *Parameterization schemes: keys to understanding numerical weather prediction models*. Cambridge University Press, Cambridge
17. Stevens B, Feingold G (2009) Untangling aerosol effects on clouds and precipitation in a buffered system. *Nature* 461:607–613
18. Taylor KE, Stouffer RJ, Meehl GA (2011) An overview of CMIP5 and the experiment design. *Bull Am Meteorol Soc*. doi:10.1175/BAMS-D-11-00094.1
19. Wang H, Rasch PJ, Feingold G (2011) Manipulating marine stratocumulus cloud amount and albedo: a process-modeling study of aerosol-cloud-precipitation interactions in response to injection of cloud condensation nuclei. *Atmos Chem Phys Discuss* 11:885–916. doi:10.5194/acpd-11-885-2011
20. Washington W, Parkinson CL (2005) *An introduction to three dimensional climate modeling*, 2nd edn. University Science, Sausalito, p 354
21. Weart SR (2010) *The discovery of global warming*. Harvard University Press, Cambridge, MA
22. Wigley TML, Raper SCB (2005) Extended scenarios for glacier melt due to anthropogenic forcing. *Geophys Res Lett* 32: L05704. doi:10.1029/2004GL021238
23. Zhang GJ, McFarlane NA (1995) Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate centre general circulation model. *Atmos Ocean* 33(3):407–446
24. Zhu P et al (2005) Intercomparison and interpretation of single-column model simulations of a nocturnal stratocumulus-topped marine boundary layer. *Mon Weather Rev* 133: 2741–2758
25. Myhre G, Highwood EJ, Shine KP, Stordal F (1998) New estimates of radiative forcing due to well mixed greenhouse gases. *Geophysical Research Letters* 25(14), pp. 2715–2718, doi:10.1029/98GL01908

---

## Avian Specific Transgenesis

MICHAEL J. MCGREW

Avian Germ Cell Laboratory, Division of Developmental Biology, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Midlothian, UK

## Article Outline

Glossary

Definition of the Subject

Introduction

Methods of Avian Transgenesis

Improved Transgenesis by the Directed Targeting of the Germ Cell Lineage

Future Directions

Bibliography

## Glossary

**Founder bird ( $F_0$ )** A bird into which a genetic vector was introduced. This vector may become incorporated into the chromosomes of some of the somatic and germ cells of the bird.

**$F_1$  bird** The offspring of the  $F_0$  bird. Transgenic  $F_1$  birds will contain the transgene incorporated into a chromosome in every cell of the animal.

**Germ line chimera** A genetically modified founder bird ( $F_0$ ) in which the introduced biological material (i.e., DNA, virus, cells) has contributed to cells of the germ cell lineage of the host animal.

**Primordial germ cells** Cells in the developing embryo belonging to the germ cell lineage at a developmental stage before any sexual differentiation has occurred. These primitive germ cells may be located in the embryo in tissues other than the forming gonad.

**Retrovirus** A virus that carries its genome as an RNA molecule. After infecting a cell, the RNA is reverse transcribed into a DNA molecule that is inserted into the genome of the infected cell at which point it is referred to as a provirus. The provirus is passed to all daughter cells as part of the host cell's genome.

**Recombinant proteins** A protein produced using recombinant DNA technology. The DNA sequence encoding the protein of interest is artificially constructed using genetic engineering and the protein is produced by inserting the DNA along with DNA regulatory regions into a bacterium, a eukaryotic cell, or an animal.

**Transgenic bird** Any avian species in which part of its genetic component contains DNA deriving from an exogenous source, or that the genome of the bird has been altered by human intervention.

**Transgenic bird lines** A flock of birds deriving from a  $F_1$  bird containing a transgenic modification at a defined genetic locus.

## Definition of the Subject

Transgenesis is the process by which an *exogenous* DNA molecule is incorporated into the genome of an animal. This technology promises the possibility to investigate and manipulate the production traits of poultry, produce recombinant proteins in the eggs of genetically engineered layer lines, and directly intervene in the health and welfare of avian species. The complexity of

the avian egg and the precocious development of the avian embryo in the female before oviposition (laying) have hindered endeavors in avian transgenesis. Three decades of effort have been carried out to achieve the genetic modification of the avian genome. The generation of novel methods for the modification of the avian genome has led to the current advances in the field of avian transgenesis. This entry will delineate the methods used for avian transgenesis, the current state of the art, and the potential future directions research in this field will take.

## Introduction

Due to its use in meat and egg production and also as a model organism for developmental biology studies in laboratories, the chicken is the most widely used bird species for the development of transgenic technologies. There are three commonly cited incentives for the development of transgenesis in the chicken. These are the production of biopharmaceutical proteins in eggs, the generation of disease-resistant flocks along with an increase in understanding disease resistance in poultry, and the generation of useful transgenic lines for studies of developmental biology. With the recent development of novel avian transgenesis techniques, these goals are beginning to be achieved.

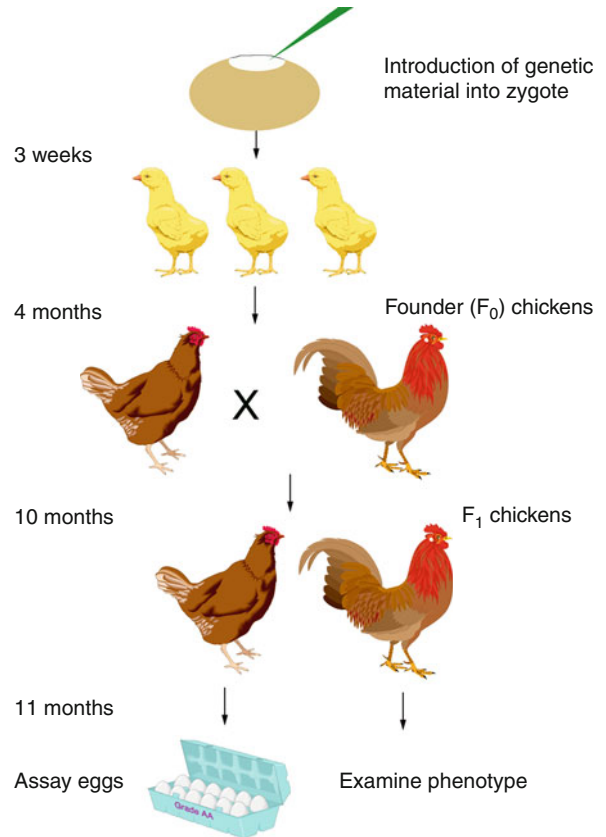
The principal rationale behind this emerging field is industry driven and is to produce a platform for the biosynthesis of biopharmaceutical proteins in chicken eggs. Animal-based bioreactors, a process sometimes referred to as animal "biopharming," were proposed as a non-cell-based method for the production of recombinant human proteins in the 1980s (reviewed in [1]). Animal-based bioreactors were inferred to be a lower cost, safe, and more efficient system in which to produce large quantities of bioactive recombinant proteins than cell-based production platforms. Efforts to produce recombinant proteins in larger animals have focused on the secretion of the target protein into the milk of the transgenic animal using the regulatory regions from a milk protein gene. Potential mammalian species commonly proposed and studied for recombinant protein production include cows, pigs, sheep, goats, lambs, and rabbits. After over two decades of investigations, the first biopharmaceutical protein produced in an animal's milk has been approved for use in

humans in Europe and the United States. This product is recombinant human antithrombin (ATryn-GTC Biotherapeutics) produced and isolated from goat's milk. Profitability in the marketplace along with consumer acceptance will be the key factor in determining the success of this and similar ventures using large animal delivery systems. In addition, the regulatory approval of this recombinant protein validates the use of animal bioreactors for protein production.

The chicken has also been proposed as a potential animal-based platform for the production of recombinant human proteins. In this case, the recombinant protein is produced in the oviduct of the hen and secreted in the forming egg as it passes down the oviduct. The target protein would be targeted to the egg white of the egg using the regulatory regions from one of several proteins expressed in the albumen of chicken eggs (reviewed in [2]). The chicken has several advantages over larger animal-based bioreactors. The chicken has a shorter generation time than a larger mammal (4 months for egg production vs. 14 months for first milk production in goats), which means a more rapid time from production of the transgenic animal until production of the recombinant protein (Fig. 1). This, in turn is reflected in the second advantage, which is the lower cost of protein production in chickens. One estimate predicts that the cost of recombinant proteins in chickens will be 1/5 of that of larger mammals [3]. This is due to the reduced costs of housing, feed, and quantity of recombinant protein produced per animal. In addition, the glycosylation pattern of secreted proteins in chickens is closer to the pattern seen on human proteins than those seen on proteins secreted in several large mammals [4].

Several studies have reported the production of bioactive recombinant protein in chicken eggs [5–7]. The methods used to produce these transgenic animals will be discussed in the following sections. These initial reports demonstrated that bioactive recombinant human proteins can be produced in the eggs of transgenic chickens. To date, excluding the production in influenza vaccines, no proteins produced in chicken eggs have been approved for use in humans.

A second reason for the production of transgenic chickens is the development of transgenic models for the study of the biology of the chicken and developmental biology in general. The chicken has long been



**Avian Specific Transgenesis. Figure 1**

Time line for production of transgenic chicken lines. The time line from the introduction of a transgenic construct into a fertilized chicken egg until the generation of  $F_1$  offspring. The  $F_1$  animals can be analyzed for gene expression or bred to transmit the transgene to the next generation

a model system for the study of developmental biology due to the accessibility of the chicken embryo in the developing egg [8]. The preeminence of the chicken as the model system for embryological studies is now taken by the mouse and the zebra fish due to the lower costs associated with maintaining colonies of these animals, quicker breeding times between generations, and most importantly, the ability to investigate gene function using transgenesis. New developments in avian transgenesis have led to the development of new avian models for developmental biology [9, 10]. The generation of transgenic lines containing the green fluorescent protein (GFP) in every cell of the



developing animal permits the fate mapping of living cells, the study of cells and their descendants during embryogenesis. This has facilitated the investigation of axial stem cell populations in the chicken embryo and the formation of a novel hypothesis for cellular sex identity in avian species [10, 11]. The development of more advanced methods of transgenesis will permit the generation of chickens containing genetically modified alleles, which will serve as models for investigations of early development, avian physiology, and production traits.

A third major reason for the generation of transgenic chickens is to use genetic modification for the increase of production traits that would lead to a decrease in environmental impact for poultry production. This could also be viewed as means of addressing the sustainability of the poultry industry with increasing demand on production and decreases in available resources. One manner in which this could be accomplished is by reducing waste production during rearing and increasing meat production for unit energy (reviewed in [12]). Several transgenic models have been developed in mammalian farm animals which exemplify the potential benefits of trait modification. A transgenic pig producing less organic phosphate waste products has been produced [13] and a similar strategy has been proposed for the chicken [14].

A comparable impact could be achieved on poultry production by reducing the losses attributable to disease. Avian transgenesis offers the potential to generate disease-resistant poultry and to investigate the genetic traits of disease resistance in poultry. It is estimated that during the avian influenza outbreak of 2004–2005, 100 million chickens were culled in Southeast Asia [15]. Using transgenic technology, the capability exists to generate a transgenic animal which will be resistant to several of the endemic diseases affecting poultry production (Avian influenza virus, Marek's disease, Newcastle disease, fowl cholera). While the production of these birds for use in meat and egg production is not currently accepted, transgenic chickens will be of use for investigating the pathogenesis of infection and the determination of viral targets of infection and pathogenesis. This in turn could generate targets for directed vaccine development. In addition, the validation of genetic traits for disease resistance also necessitates an analysis of the disease loci using transgenic technology

in avian cells or the whole animal. Currently, no disease-resistant transgenic chicken models have been developed, but the production of transgenic cattle with resistance to mastitis suggests that this will be a profitable area of investigation [16]. With increasing demands on poultry production, an increase in methods of disease resistance will be needed.

## Methods of Avian Transgenesis

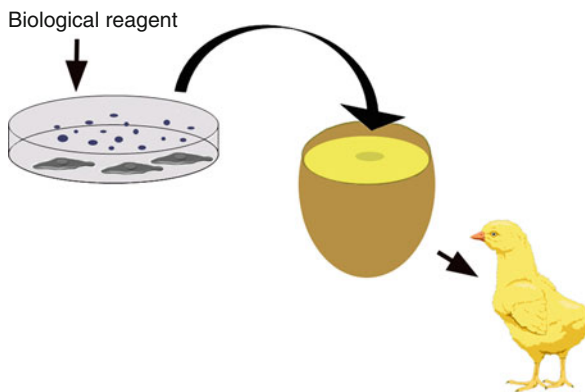
The first efficient method developed for the introduction of exogenous DNA into mammals to create germ line chimeras was using pronuclear injection [17]. This technique entails the microinjection of DNA encoding the gene of interest (transgene) into the male pronucleus of the recently fertilized oocyte at a developmental stage previous to the first cell cleavage. Although pronuclear microinjection was primarily developed for transgenesis in the mouse, it has been exploited to generate transgenic animals in rats, sheep, cows, pigs, and rabbits [18–20]. Rates of transgenesis using pronuclear injection (transfer of the injected transgene to the genome of the offspring) in species other than mice are low. In general, only 1–5% of the offspring will have incorporated the transgene into their genome.

The use of pronuclear injection is even more problematic in the chicken. The laid chicken egg consists of a single epithelial layer of unpatterned cells termed the blastoderm consisting of 45,000–60,000 cells [21]. Injection of any biological material at this stage signifies that only a small proportion of the embryonic cells will be potentially exposed to and incorporate the transgene into their genome. The development of the chicken zygote in the hen before the egg is laid necessitates that pronuclear microinjection can only be carried out with surgical removal of the fertilized oocyte from the hen. This entails that the egg is subsequently surgically transferred to a surrogate host hen or cultured *in vitro* for 21 days until hatching [22]. The ovum of the avian egg is the large opaque yolk of the egg. The ovum contains a small pool of cytoplasm lying on the yolk material in which the pronuclei are located. Microinjection into the pronuclei is not possible because of the opacity of the yolk material underlying the nuclei; thus any injected biological material must be injected into the cytoplasm surrounding the pronuclei. For this reason, the rate of transgenesis for using

pronuclear injection chickens is much lower than in other vertebrates and the delivery of biological material is usually performed in the laid egg (Fig. 2).

### Injection of DNA into Early Avian Oocytes

Microinjection of DNA transgenic constructs into the early chicken oocyte has proven to be a feasible procedure for the production of transgenic chickens. As described above, the DNA construct is injected into the cytoplasm of the fertilized oocyte at an early developmental stage (Fig. 2). The egg is subsequently cultured *ex ovo* in a host shell until normal hatching occurs. Love [23] introduced a  $\beta$ -galactosidase reporter construct into the fertilized chicken oocyte and cultured the embryos *ex ovo* until hatching. In this report, 50% of the injected embryos contained the transgene and 14% of the hatched founder birds transmitted the transgene to the next generation. This is a similar transmission rate to that observed in mammalian species using microinjection. A comparable result was observed in which founder transgenic chickens were obtained using microinjection and a modified *ex ovo*



### Avian Specific Transgenesis. Figure 2

Introduction of genetic material into the fertilized zygote. The genetic material to be introduced into the egg (DNA, viruses, transposons, purified proteins) is either first transfected into cells for packaging or directly introduced into the egg at an early stage of development (laid egg or fertilized oocyte). The embryo is subsequently incubated until hatching in its own shell or a host chicken shell. The hatchling will then be analyzed for the genetic modification

culture system [24]. In this example, no germ line transmission to the  $F_1$  generation was observed. Although it is possible to produce a transgenic bird using microinjection, this avenue of research has not proven fruitful due to the cost and difficulty in obtaining and manipulating the early chicken oocyte. Instead, research has focused the use of retroviral vectors for the introduction of transgenes into the genome of the chicken at laid egg stages of development.

### Replication-Competent Viral Vectors

A more efficient method to create genetically modified chickens is using retroviral vectors to introduce transgenes into the chicken genome. Retroviruses are RNA viruses that reverse transcribe the RNA genome into a DNA intermediary that is integrated into the genome of the infected cell. The integrated virus, or provirus, transcribes new copies of the viral genome that are packaged into new viral particles that bud from the infected cell and in turn infect neighboring cells. In this manner, most cells of the developing embryo will be infected along with the germ cells. Germ cells containing the integrated provirus in their genome will transmit this modification to all of the cells of the offspring deriving from that germ cell. Using recombinant DNA technology, it is possible to remove part of the viral genome and replace it with an inserted transgene. In this manner, a line of genetically modified chickens can be produced containing a transgene of interest.

The first use of retroviruses in chicken for transgenesis used viruses that were replication-competent (i.e., infective). In this case, the virus was injected at the laid egg stage and incubated until hatching. Though initially only a small portion of the blastoderm may be infected, because the viral infection spreads during development, many more cells will contain proviral inserts. The first demonstration of the genetic modification of the chicken germ line was accomplished using a replication-competent avian retrovirus. Salter [25] injected either reticuloendothelial virus or avian leukosis virus under the blastoderm of laid eggs. They demonstrated that the integrated provirus could be found in the offspring of the injected hosts [25]. As a further demonstration of this technology, using the ALV retrovirus, 24% of founder animals

contained the integrated provirus and this was transmitted to  $F_1$  offspring with a transmission rate between 1% and 11% [26]. A replication-competent retrovirus was also used to introduce a transgene containing the bovine growth hormone under control of the mouse metallothionein promoter into founder ( $F_0$ ) chicken [27]. Thus, replication-competent viruses offered an initial alternative to pronuclear injection for chicken transgenesis.

It must be noted that replication-competent viruses are limited in their usefulness in transgenesis for several reasons. First, since replicative virus can reintegrate into the genome, it is difficult to correlate a single integration site with the synthesis of recombinant protein from the integrated transgene. This is important as any regulatory approval procedure will require genetically defined transgenic animals containing a single defined transgenic insert which produces consistent levels of recombinant protein [28]. Second, insertion mutagenesis caused by the insertion of a retrovirus can disrupt an endogenous genetic locus and also lead to malignant cell growth in the host animal (reviewed in [29]). Third, and most important, a process called viral or transgene silencing is often observed in animals containing replication-competent proviral inserts. In transgene silencing, the transgene becomes silenced after it is introduced into early embryos, which is correlated with a high level of cytosine methylation throughout the proviral genome [30]. In the transgenic offspring of these founder animals, the transgene is not expressed or expression levels are severely reduced and the proviral genome is usually methylated in the long terminal repeat (LTR) region. It is believed that retroviral silencing occurs in the stem cell populations of the early embryo and leads to permanent silencing of transcription from the integrated provirus (reviewed in [31]). For these aforementioned reasons, replication-defective retroviruses have become the tool of choice for avian transgenesis.

### Replication-Defective Viruses

Replication-defective retroviruses are modified retroviral vectors that are capable of infecting a host cell but lack key viral genes necessary to complete the viral life cycle and produce infective viral particles. In replication-defective retroviral vectors, a large portion of the

viral genome has been removed from the virus and these genes are supplied in *trans* to allow for packaging of the defective viral particles. The envelope proteins of the viral vector can also be modified (viral pseudo-typing) to permit infection of cell types and species normally outside of the retrovirus host range. The integrated provirus will contain a viral promoter in the LTR and few viral genes. It is also possible to incorporate a large transgene (8–9 kb) into the provirus in place of the truncated viral genome.

The first demonstration of germ line transgenesis in chicken used a replication-defective reticuloendothelial retroviral vector containing a neomycin reporter gene injected into the blastoderm of laid eggs [32]. Twenty-two percent of the  $F_0$  hatchlings contained proviral integrations and the transmission rate to the  $F_1$  generation was between 2–8%. Expression of the neomycin transgene in these birds was not demonstrated.

Since this first report, replication-defective retroviruses have been used extensively for chicken transgenesis. The most common retroviral vectors used are spleen necrosis virus, avian leukosis virus, reticuloendothelial virus mentioned previously, and Moloney murine leukemia virus (MMLV). All of these vectors have been used to generate several lines of transgenic chickens [33–36]. Transmission rates of the integrated provirus after blastodermal injection from  $F_0$  cockerels to  $F_1$  offspring have been reported to be as high as 8%.

A novel development of this technology is to induce expression of the recombinant protein at defined timepoints to avoid possible transgene toxicity during development. This was achieved by using a replication-defective MMLV containing a tetracycline-regulated promoter to drive expression of human-recombinant erythropoietin protein in chicken eggs only in the presence of the inducer, doxycycline [7]. More than 90% of the  $F_0$  chickens contained proviral integrations, and these birds transmitted the transgene to between 0.7% and 1.8% of the  $F_1$  generation. Recombinant erythropoietin was produced in eggs after doxycycline induction.

In some replication-defective retroviral vectors, there is ongoing transcription from viral regulatory LTRs in the proviral insert. This transcription is thought to lead to a partial silencing of the adjacent transgene. An example of transgene silencing in

birds was observed in transgenic quail containing a replication-defective MMLV using a LTR to drive expression of a GFP reporter construct [37]. Transgene expression was silenced in F<sub>1</sub> and F<sub>2</sub> birds. Other lines of transgenic chicken have displayed some suggestions of silencing between generations using the MMLV virus [38]. Researchers have attempted to circumvent this problem by using modified MMLV vectors containing internal promoters to drive transgene expression. This strategy appears to have worked in some transgenic birds that had no transgene silencing between generations [39]. Still, because of the perceived problems of silencing of these retroviral vectors, researchers have turned to a different class of retroviruses, the lentiviruses, for avian transgenesis.

### Lentiviral Vectors

Lentiviral vectors have primarily been developed as potential viral agents for gene therapy in humans. Some of the advantages of lentiviral vectors are that the integrated provirus is preferentially incorporated into the open chromatin structure surrounding expressed genes, lentivectors will infect and integrate into the genome of post-mitotic cells, and the viral vector accepts a transgenic insert up to 9 kb in size. Moreover, there appears to be no transgene silencing upon transmission of the integrated transgene between generations [40, 41].

As an alternative to other retroviral vectors and to pronuclear microinjection, replication-defective retroviruses of the lentivirus class have proven to be very efficient for use in mammalian transgenesis [42, 43]. These vectors have been used successfully to generate transgenic cattle, pigs, rats, sheep, rabbits, and primates [44–48]. In these cases, the viral particles are injected into the perivitelline space surrounding the fertilized zygote. The zygote is reintroduced into surrogate hosts and the animals are bred and tested for germ line transmission.

The first demonstration of the use of lentiviral vectors for avian transgenesis used viral vectors containing transgenes encoding either GFP or  $\beta$ -galactosidase reporter genes [49]. Viral particles were injected into the blastoderm of laid eggs as described above. Efficiency of transmission of the integrated transgene from founders to the F<sub>1</sub> generation

was between 4% and 45% and no transgene silencing was observed between generations. Since this report, lentiviral vectors have been used successfully to generate transgenic chickens containing a range of transgenes [6, 9, 10, 50, 51] and transgenic quail [52–54] and also transgenic zebra finch [55]. These vectors have also proven useful to generate biopharmaceutical proteins in eggs. Recombinant human  $\beta$ -interferon has been produced in the egg white of transgenic chicken [6], and recombinant human interleukin-1 receptor antagonist has also been produced in the eggs of quail [56].

While lentiviral vectors may now be the preferred method for generating transgenic birds, there will be public concerns with the use of transgenic birds containing retroviral transgenes, in particular, HIV-based lentiviral transgenes, for the production of biopharmaceutical recombinant proteins and the development of advanced production traits in avian species. It remains to be determined if consumer acceptance of these products will be forthcoming.

### Transposons

Transposable elements were first identified by the seminal work by Barbara McClintock in maize [57]. Transposable elements are mobile genetic elements found in the genome of all organisms and are able to move (i.e., translocate) from one region of the genome to another. These are modules of selfish DNA that exist only to replicate their own DNA and are postulated as parasitic invaders of most genomes [58]. Transposable elements differ from viruses in that they do not spread by infection of neighboring cells; they are usually passed passively through cell division to daughter cells. Class II DNA transposons are a distinct type of transposable element that moves through the genome using a “cut and paste” method. The DNA transposon encodes an enzyme called a transposase which mediates the removal of the transposon from one chromosomal location and the insertion into a second chromosomal location. DNA transposons have been modified for the use in transgenesis in many vertebrates. This entails that the coding sequence for the transposase is deleted from the transposon and subsequently introduced into the cell with the transposon in *trans* [59]. The transposon will be incorporated into the genome of the cell but

cannot “jump” to a new genomic location because it lacks the transposase gene. This is similar in approach to the packaging of replication-defective retroviruses for transgene delivery. The truncated transposon can be engineered to contain additional DNA sequences which will be inserted into the genome of the cell upon transposon integration. The size of the transgene carried by the transposon can be up to 10 kb in size [60, 61].

Transposable elements have been shown to be efficient in the generation of transgenic chickens. The mariner transposon from *Drosophila maritiana* containing an internal transposase was used to generate transgenic chickens [62]. Mariner was microinjected into the fertilized zygote and chickens were hatched and bred to generate F<sub>1</sub> birds. One founder bird transmitted the mariner transposon to 29% of its offspring. This result demonstrates that transposable elements can be used to modify the genome of birds although no transgene cargo was introduced in this example. So far, no reports have been made to generating a transgenic chicken line using transposons containing a transgene. However, the Tol2 and piggyBac transposons have been shown to be functional in transient transgenesis in the chicken [63, 64] and other transposable elements have been used successfully for mammalian transgenesis [59].

### Use of Chicken Embryonic Stem Cells

Whilst the early embryonic development of the chicken is significantly different from that of a mammal, cells of the early chicken blastoderm do contribute to somatic and germ cell lineages when transferred to host chicken embryos [65]. This indicates that the early chicken blastoderm may have the developmental equivalence of the inner cell mass of the early mouse embryo, further suggesting that embryonic cell lines derived from the chicken blastoderm may be developmentally equivalent to mouse embryonic stem cells. In support of this hypothesis, several groups were able to derive cell lines from early chicken embryos that were putatively identified as chicken embryonic stem cells because of their ability to contribute to many tissues in chimeric birds [66, 67]. The isolation of embryonic cell lines that contribute to the forming avian embryo offers the opportunity to carry out gene targeting,

whereby the introduced transgene can be targeted to a precise genetic locus. Gene targeting uses the process of homologous recombination. In this process, one genomic copy of a gene can be replaced by a second copy of this gene using normal DNA repair mechanisms. The replacement gene will display homology over most of its length but can be engineered to contain a foreign piece of DNA which could encode a recombinant protein or a selection marker. Using homologous recombination, researchers can disrupt a genetic locus, introduce a gene encoding a recombinant protein, or replace one genetic allele with a second allele. Homologous recombination has been carried out in chicken cells [68]. Genetic modification of chicken embryonic stem cells also has been achieved and chimeric animals have been produced which contain extensive contribution of these cells to all germ layers of the forming embryo [69].

An important caveat to this work is to note here that the germ cell lineage is segregated from the somatic cell lineage very early in development (see section “[Methods of Avian Transgenesis](#)”). It is believed that for this reason cESCs do not contribute to the germ lineage in chimeras after more than 7 days of in vitro culture [67, 69]. So, although cESCs may be useful for creating somatic chimeras, these cells will not be useful for generating lines of transgenic chickens.

Recent experimental evidence suggests that it may be possible to differentiate cESCs into germ cells in vitro. These cells could then be introduced into host birds and the cells could differentiate into gametes and be transmitted to the next generation [70]. If this is achievable, it will create the possibility to carry out gene targeting in the chicken and create transgenic chicken lines using homologous recombination in chicken embryonic stem cells.

### Improved Transgenesis by the Directed Targeting of the Germ Cell Lineage

Most attempts to generate transgenic chickens have resulted in inefficient transmission of the integrated transgene from F<sub>0</sub> to F<sub>1</sub> animals. This inefficiency leads to an increase in cost for production of transgenic animals due to increase in both time and breeding numbers of F<sub>0</sub> founder birds (reviewed in [2]). Several methods have been developed to increase the rate of



transgenesis. These methods attempt to amplify the interaction of the exogenously introduced DNA or virus with the germ cell lineage of the avian embryo. The avian germ cell lineage comprises the only cells of the animal that will contribute genetic material to the next generation. Two methods developed to increase the interaction of DNA/virus with the avian germ cells are described in the next section. These methods are the purification of primordial germ cells from the developing embryo and the propagation of primordial germ cells for use as a cell-based method of transgenesis.

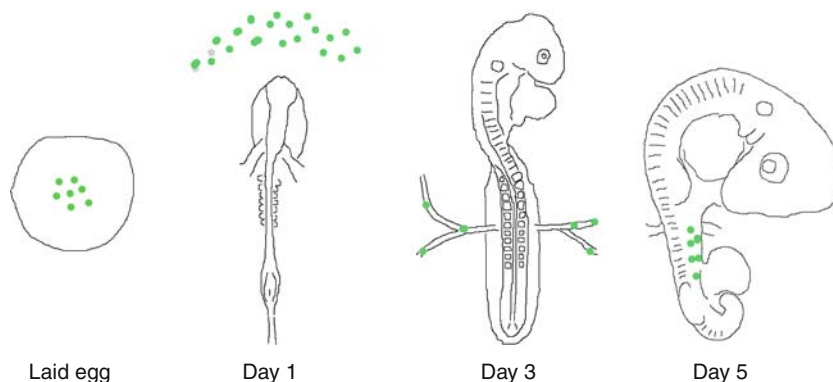
The germ cell lineage in birds is considered to be determined. The maternal determinants (protein, RNA) that specify the germ cell fate are thought to be deposited in the developing oocyte as it matures in the hen. During the initial segmentation of the zygote, these factors (maternal determinants) are segregated into a small number of cells of the forming embryo [71]. Descendants of these early germ cells will give rise to all cells of the germ cell lineage and are referred to as primitive or “primordial” germ cells. Moreover, since the germ cell lineage is segregated from the somatic cell lineage so precociously, chicken embryonic stem cells derived from the blastoderm will not contribute to the germ lineage when reintroduced into early embryos. The primordial germ cells are initially found in the center of the blastoderm of the laid egg [72] (Fig. 3). These cells actively migrate anteriorly in the embryo to an extraembryonic location near the future head, termed the germinal crescent. When the primitive

circulatory system forms, the primordial germ cells enter the circulation and are carried to the lateral plate mesoderm adjoining the prospective gonad. During the next 2 days of development, the primordial germ cells migrate through the lateral plate mesoderm and enter the developing gonad. The germ cells subsequently differentiate into the gametes, sperm and eggs, in the maturing gonad. In the hens, the germ cells enter meiosis by day 16 of development. In the adult cockerel, a stem cell population of germ cells remain, the spermatogonial stem cells, which generate spermatogonia during the life of the bird.

All methods of germ line transgenesis in all species are dependent on the integration of the introduced transgene into the genome of a germ cell, which will subsequently contribute its DNA to the next generation. Any interactions between the exogenously introduced DNA/virus and non-germ-line cells of the animal will lead to a reduction of transmission of the transgene to the F<sub>1</sub> birds. Thus, investigations in the chicken have concentrated on increasing the interaction of the transgenic vector with the germ cells of the founder animal to increase the rate of transgenesis. The next sections detail methods to isolate avian germ cells and increase this interaction.

### Germ Cell Purification

The germ cell lineage of birds can be isolated and returned to a host embryo in which these donor germ



**Avian Specific Transgenesis. Figure 3**

Germ cell migration in the developing embryo. The primordial germ cells are initially found in the center of the developing blastoderm. From here the cells migrate anterior to the head region. The cells enter the circulation and congregate in the lateral plate adjoining the future gonad. The primordial germ cells subsequently migrate into the forming gonad

cells will migrate to the gonad and produce functional gametes [73, 74]. The intrinsic ability of donor germ cells to colonize the host germ cell lineage extends to germ cells from the germinal crescent, the circulatory blood, or post-migratory germ cells in the nascent gonad (Fig. 3). This provides the opportunity to genetically modify the donor germ cells before their introduction into host embryos. The first demonstration of this technique was carried out by Vick [75]. Primordial germ cells from the blood or the germinal crescent were incubated with replication-defective avian leukosis virus and subsequently returned to a host embryo. They observed that 3–23% of founder chickens were transgenic, and these birds transmitted the transgene to 2–4% of F<sub>1</sub> birds.

Several methods have been developed to purify PGCs from the developing embryo before which they are subsequently incubated with virus or transfected to introduce a transgenic vector before being reintroduced into host embryos. Primordial germ cells express the pluripotent stem cell marker, SSEA-1. Using an antibody to SSEA-1, these cells can be purified by either magnetic-activated cell sorting (MACS) or fluorescence-activated cell sorting (FACS). Both isolation techniques have been optimized for the isolation of thousands of germ cells from the blood or embryonic gonads of chicken embryos. After short periods of incubation, the isolated primordial germ cells still retain the ability to migrate to the gonad and contribute to the germ cell lineage [76]. Similarly, due to the large size of the PGC in comparison to other cells in the blood, it is possible to isolate these cells from the blood using density gradients. Density gradients have been used successfully for the purification of PGCs [77–80].

Kim [49] isolated PGCs from day 5 gonads using MACS. These cells were incubated with a lentivirus for 6 h and reinjected in host embryos. One (of 21) injected cockerels transmitted the transgene to F<sub>1</sub> offspring. A similar approach was used to generate transgenic quail. The rate of transmission from founder to F<sub>1</sub> birds was 1.9%, which was almost equivalent to what was found when the virus was injected under the blastoderm (1.6%) [54]. These results suggest that PGC isolation did not increase rates of transgenesis in comparison to blastoderm injection but further improvements on the interaction of the virus with the germ cells may augment rates of transmission.

## Propagation of the Avian Germ Cell Lineage

The *in vitro* propagation of germ cells offers the prospect to carry out gene targeting in avian species. Both spermatogonial stem cells [81] and gonadal embryonic germ cells [82] have been cultured *in vitro* for shorter periods (5 and 30 days, respectively) and after culturing have contributed to the germ lineage. Moreover, the cultured gonadal embryonic germ cells generated F<sub>1</sub> offspring. These culture techniques, if extended, could offer the opportunity to carry out gene targeting in chickens. Recently, a breakthrough was described for the culture of primordial germ cells from the circulatory phase of development (Fig. 3). The culture of embryonic blood from day 3 chicken embryos using modified avian embryonic stem cells protocols led to the long-term expansion of primordial germ cells *in vitro* [83]. These cells contained telomerase activity suggesting that they may be immortal and the transmission of genetic material from both the male and female germ cells to subsequent generations was demonstrated after 110 days in culture (0.1–86% transmission). The authors genetically modified the cultured primordial germ cells using a transgene flanked by insulator sequences to prevent silencing. Founder birds containing the modified germ cells transmitted the transgene to 1–92% of F<sub>1</sub> birds which accurately expressed the GFP transgene. Important caveats for this procedure were that the derivation efficiency of primordial germ cell lines was low (12%), and this culture technique has not been repeated by other research groups. It was later shown that gene targeting using homologous recombination is also possible in cultured primordial germ cells [84]. This research presents the prospect of carrying out gene targeting in avian species for the production of pharmaceutical recombinant proteins and to investigate the function of genetic pathways that may be involved in production traits by directly modifying these pathways.

## Future Directions

The field of avian transgenesis has progressed rapidly in the last 5 years. Useful transgenic chicken models for use in biological studies have been developed [9, 10, 51, 85]. In the near future, it can be expected that a plethora of new transgenic models for both

production traits and disease resistance to be generated. These transgenic models will generate impetus for trait selection breeding programs and rational vaccine development, first and foremost in the investigation of avian influenza.

Future uncertainties exist regarding the potential for commercialization of biopharmaceuticals proteins produced in chicken eggs and the acceptance of genetically modified chickens in general. To date, several recombinant human proteins have been produced in chicken eggs at levels up to 1 mg of protein per egg [5–7, 38, 56]. In addition, it remains to be seen if there will be consumer acceptance of genetically modified poultry products and if a functional transgenic chicken product will be brought to “market.” The benefits of increased disease resistance and increased feed conversion will need to be weighed against consumer demand and acceptance. These opposing forces will have direct impact on the sustainability of the poultry industry as food demands continue to grow. The issuing of guidelines by the FDA for the regulation of genetically modified animals containing recombinant DNA should facilitate this process [28].

The development of new vectors, both viral and nonviral, for transgenesis should decrease the cost and accelerate the development of new avian transgenic models and the use of nonviral vectors may assuage public fears of transgenic modification of farm livestock. DNA transposons hold great promise as useful vectors for avian transgenesis. The next few years should bring new advances as some of the DNA transposable elements that are showing promise in mammalian transgenesis (Tol2, PiggyBac, Sleeping Beauty) are used for chicken transgenesis.

The greatest advance in avian transgenesis lies in the development of efficient gene targeting of specific genomic loci using homologous recombination. This objective may be accomplished using SSCs, gonadal embryonic germ cells, or primordial germ cells [80–83]. Gene targeting of primordial germ cells appears to be a viable technique [84] although it remains to be proven if this will be a viable method for avian transgenesis. The use of zinc finger nucleases which target specific loci in the vertebrate genome has proven to be useful for gene targeting in both zebra fish and rats [85, 86]. These proteins hold great promise for the potential manipulation of the

avian genome. The generation of induced pluripotent cells (IPS cells) from somatic cells also needs to be investigated in avian species [70]. If this technique were further coupled with the conversion to germ cells, it would lay open new possibilities in avian cloning for species rescue.

Primordial germ cell culture also promises to be a viable technology for the preservation of avian germplasm (reviewed in [87]). Semen cryopreservation in avian species has proven problematic and for this reason an alternative method of germplasm cryopreservation is needed. An efficient method to preserve avian genetic resources is of importance for commercial breeders and also rare breed conservationists. The development of a robust method for avian cryopreservation would significantly enhance the biosecurity of the avian breeding industry and greatly reduce animal numbers bred for line maintenance. The use of PGCs could provide an alternative source of cellular material for cryopreservation efforts. At this stage, PGC cryopreservation offers only a slight advantage over semen cryopreservation for the preservation of single traits. Therefore, what is needed to extend this technology is to develop improved culture conditions for the efficient expansion and cryopreservation of both male and female PGCs from different breeds of chickens. This technology will need to be coupled with an efficient methodology to reconstitute the *complete genome* of avian lines from the cryopreserved cells.

Several recent reports have described transgenic offspring from quail [52, 54] and zebra finches [55] using the genetic techniques described above. To date, no transgenic modifications have been published for turkeys and duck. Since these are also important food production animals, it can be expected that advances in transgenesis will be made in these species in the near future. These potential models will be significant to the study of avian production traits but may also be pertinent to the investigation of avian influenza in which natural disease reservoirs reside in wild duck populations. It will be difficult to foresee if the transgenic technology being developed for the chicken will be easily transferred to other poultry species beyond the use of retroviral vectors. An illustration of foreseeable problems comes from investigations of embryonic stem cells. Derivation of ES cells from mammals

excluding the mouse has proven to be exceedingly difficult to achieve [88]. This may also hold true for avian transgenesis.

## Bibliography

### Primary Literature

- Houdebine L-M (2009) Production of pharmaceutical proteins by transgenic animals. *Comp Immunol Microbiol Infect Dis* 32:2107–2121
- Lillico SG, McGrew MJ, Sherman A, Sang HM (2005) Transgenic chickens as bioreactors for protein-based drugs. *Drug Discov Today* 10:3191–3196
- Council for Agricultural Science and Technology (CAST) (2007) The role of transgenic livestock in the treatment of human disease. Issue Paper 35. CAST, Ames
- Raju TS, Briggs JB, Borge SM, Jones AJS (2000) Species-specific variation in glycosylation of IgG: evidence for the species-specific sialylation and branch-specific galactosylation and importance for engineering recombinant glycoprotein therapeutics. *Glycobiology* 10:477–486
- Rapp JC, Harvey AJ, Speksnijder GL, Hu W, Ivarie R (2003) Biologically active human interferon alpha-2b produced in the egg white of transgenic hens. *Transgenic Res* 12:569–575
- Lillico SG, Sherman A, McGrew MJ, Robertson CD, Smith J, Haslam C, Barnard P, Radcliffe PA, Mitrophanous KA, Elliot EA, Sang HM (2007) Oviduct-specific expression of two therapeutic proteins in transgenic hens. *Proc Natl Acad Sci USA* 104:1771–1776
- Koo BC, Kwon MS, Lee H, Kim M, Kim D, Roh JY, Park YY, Cui XS, Kim NH, Byun SJ, Kim T (2010) Tetracycline-dependent expression of the human erythropoietin gene in transgenic chickens. *Transgenic Res* 19(3):437–447
- Stern CD (2005) The chick: a great model system becomes even greater. *Dev Cell* 8:9–17
- Chapman SC, Lawson A, Macarthur WC, Wiese RJ, Loechel RH, Burgos-Trinidad M, Wakefield JK, Ramabhadran R, Mauch TJ, Schoenwolf GC (2005) Ubiquitous GFP expression in transgenic chickens using a lentiviral vector. *Development* 132:935–940
- McGrew MJ, Sherman A, Lillico SG, Ellard FM, Radcliffe PA, Gilhooley HJ, Mitrophanous KA, Cambray N, Wilson V, Sang H (2008) Localised axial progenitor cell populations in the avian tail bud are not committed to a posterior Hox identity. *Development* 135:2289–2299
- Zhao D, McBride D, Nandi S, McQueen HA, McGrew MJ, Hocking PM, Lewis PD, Sang HM, Clinton M (2010) Somatic sex identity is cell-autonomous in the chicken. *Nature* 464:237–242
- Council for Agricultural Science and Technology (CAST) (2009) Animal productivity and genetic diversity: cloned and transgenic animals. Issue Paper 43. CAST, Ames
- Golovan SP, Meidinger RG, Ajakaiye A, Cottrill M, Wiederkehr MZ, Barney DJ, Plante C, Pollard DW, Fan MZ, Hayes MA, Laursen J, Hjorth JP, Hacker RR, Phillips JP, Forsberg CW (2001) Pigs expressing salivary phytase produce low-phosphorus manure. *Nat Biotechnol* 19:741–745
- Cho J, Choi K, Darden T, Reynolds PR, Pettite JN, Shears SB (2006) Avian multiple inositol polyphosphate phosphatase is an active phytase that can be engineered to help ameliorate the planet's "phosphate crisis". *J Biotechnol* 126:248–259
- Tiensin T, Chaitaweesub P, Songserm T, Chaisingh A, Hoonsuan W et al (2005) Highly pathogenic avian influenza H5N1, Thailand, 2004. *Emerg Infect Dis* 11:1664–1672
- Wall RJ, Powell A, Paape MJ, Kerr DA, Bannermann SS, Pursel CG, Well KD, Talbot N, Hawk HW (2005) Genetically enhanced cows resist intramammary *Staphylococcus aureus* infection. *Nat Biotechnol* 23:445–451
- Gordon JW, Scangos GA, Plotkin DJ, Barbosa JA, Ruddle FH (1980) Genetic transformation of mouse embryos by microinjection of purified DNA. *Proc Natl Acad Sci USA* 77:7380–7384
- Hammer RE, Pursel VG, Rexroad CE, Wall RJ, Bolt DJ, Ebert KM, Palmiter RD, Brinster RL (1985) Production of transgenic rabbits, sheep and pigs by microinjection. *Nature* 315:680–683
- Mullins JJ, Peters J, Ganten D (1990) Fulminant hypertension in transgenic rats harbouring the mouse Ren-2 gene. *Nature* 344:541–544
- Krimpenfort P, Rademakers A, Eyestone W, van der Schans A, van den Broek S, Kooiman P, Kootwijk E, Platenburg G, Pieper F, Strijker R, de Boer H (1991) Generation of transgenic dairy cattle using 'in vitro' embryo production. *Biotechnology (N Y)* 9:844–847
- Kochav S, Ginsburg M, Eyal-Giladi H (1980) From cleavage to primitive streak formation: a complementary normal table and a new look at the first stages of the development of the chick. II. Microscopic anatomy and cell population dynamics. *Dev Biol* 79:296–308
- Perry MM (1988) A complete culture system for the chick embryo. *Nature* 331:70–72
- Love J, Gribbin C, Mather C, Sang H (1994) Transgenic birds by DNA microinjection. *Biotechnology* 12:60–63
- Naito M, Sasaki E, Ohtaki M, Sakurai M (1994) Introduction of exogenous DNA into somatic and germ cells of chickens by microinjection into the germinal disc of fertilized ova. *Mol Reprod Dev* 37:167–171
- Salter DW, Smith EJ, Hughes SH, Wright SE, Fadly AM, Witter RL, Crittenden LB (1986) Gene insertion into the chicken germ line by retroviruses. *Poult Sci* 65:1445–1458
- Salter DW, Smith EJ, Hughes SH, Wright SE, Crittenden LB (1987) Transgenic chickens: insertion of retroviral genes into the chicken germ line. *Virology* 157:236–240
- Chen HY, Garber EA, Mills E, Smith J, Kopchick JJ, DiLella AG, Smith RG (1990) Vectors, promoters, and expression of genes in chick embryos. *J Reprod Fertil Suppl* 41:173–1782
- U.S. Food and Drug Administration (2009) Guidance for industry #187: regulation of genetically engineered animals containing heritable recombinant DNA constructs. Final Guidance
- Romano G, Marino IR, Pentimalli F, Adamo V, Giordano A (2009) Insertional mutagenesis and development of malignancies induced by integrating gene delivery

- systems: implications for the design of safer gene-based interventions in patients. *Drug News Perspect* 22:185–196
30. Jahner D, Stuhlmann H, Stewart CL, Harbers K, Lohler J, Simon I, Jaenisch R (1982) De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature* 298:623–628
  31. Pannell D, Ellis J (2001) Silencing of gene expression: implications for design of retrovirus vectors. *Rev Med Virol* 11: 205–217
  32. Bosselman RA, Hsu R-Y, Boggs T, Hu S, Bruszewski J, Ou S, Kozar L, Martin F, Green C, Jacobsen F, Nicolson M, Schultz JA, Semon KM, Rishell W, Stewart RG (1989) Germline transmission of exogenous genes in the chicken. *Science* 243:533–535
  33. Thoraval P, Afanassieff M, Cosset FL, Lasserre F, Verdier G, Coudert F, Dambrine G (1995) Germline transmission of exogenous genes in chickens using helper-free ecotropic avian leukosis virus-based vectors. *Transgenic Res* 4:369–377
  34. Harvey AJ, Speksnijder G, Baugh LR, Morris JA, Ivarie R (2002) Expression of exogenous protein in the egg white of transgenic chickens. *Nat Biotechnol* 20:396–399
  35. Mozdziak PE, Borwornpinyo S, McCoy DW, Petitte JN (2003) Development of transgenic chickens expressing bacterial beta-galactosidase. *Dev Dyn* 226:439–445
  36. Kwon MS, Koo BC, Choi BR, Park YY, Lee YM, Suh HS, Park YS, Lee HT, Kim JH, Roh JY, Kim NH, Kim T (2008) Generation of transgenic chickens that produce bioactive human granulocyte-colony stimulating factor. *Mol Reprod Dev* 75:1120–1126
  37. Mizuarai S, Ono K, Yamaguchi K, Nishijima K, Kamiyama M, Iijima S (2001) Production of transgenic quails with high frequency of germline transmission using VSV-G pseudotyped retroviral vector. *Biochem Biophys Res Commun* 286:456–463
  38. Kamiyama M, Ono K, Esaka K, Nishijima K, Kigaku R, Komatsu H, Yamashita T, Kyogoku K, Iijima S (2005) High-level expression of single-chain Fv-Fc fusion protein in serum and egg white of genetically manipulated chickens by using a retroviral vector. *J Virol* 79:10864–10874
  39. Koo BC, Kwon MS, Choi BR, Kim JH, Cho SK, Sohn SH, Cho EJ, Lee HT, Chang W, Jeon I, Park JK, Park JB, Kim T (2006) Production of germline transgenic chickens expressing enhanced green fluorescent protein using a MoMLV-based retrovirus vector. *FASEB J* 20:2251–2260
  40. Cockrell AS, Kafri T (2007) Gene delivery by lentivirus vectors. *Mol Biotechnol* 36:184–204
  41. Wang Y, Song Y-T, Liu Q, Liu C, Wang L-L, Liu Y, Zhou X-Y, Wu J, Wei H (2010) Quantitative analysis of lentiviral transgene expression in mice over seven generations. *Transgenic Res* 19:775–784
  42. Lois C, Hong EJ, Pease S, Brown EJ, Baltimore D (2002) Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science* 295:868–872
  43. Pfeifer A, Ikawa M, Dayn Y, Verma IM (2002) Transgenesis by lentiviral vectors: lack of gene silencing in mammalian embryonic stem cells and preimplantation embryos. *Proc Natl Acad Sci USA* 99:2140–2145
  44. Hamra FK, Gatlin J, Chapman KM, Grellhesl DM, Garcia JV, Hammer RE, Garbers DL (2002) Production of transgenic rats by lentiviral transduction of male germ-line stem cells. *Proc Natl Acad Sci USA* 99:14931–14936
  45. Hofmann A, Kessler B, Ewerling S, Weppert M, Vogg B, Ludwig H, Stojkovic M, Boelhaue M, Brem G, Wolf E, Pfeifer A (2003) Efficient transgenesis in farm animals by lentiviral vectors. *EMBO Rep* 4:1054–1060
  46. Ritchie WA, King T, Neil C, Carlisle AJ, Lillico S, McLachlan G, Whitelaw CB (2009) Transgenic sheep designed for transplantation studies. *Mol Reprod Dev* 76:61–64
  47. Hiripi L, Negre D, Cosset FL, Kvell K, Czömpöly T, Baranyi M, Gócza E, Hoffmann O, Bender B, Bösze Z (2010) Transgenic rabbit production with simian immunodeficiency virus-derived lentiviral vector. *Transgenic Res* Epub:Jan 13
  48. Sasaki E, Suemizu H, Shimada A, Hanazawa K, Oiwa R, Kamioka M, Tomioka I, Sotomaru Y, Hirakawa R, Eto T, Shiozawa S, Maeda T, Ito M, Ito R, Kito C, Yagihashi C, Kawai K, Miyoshi H, Tanioka Y, Tamaoki N, Habu S, Okano H, Nomura T (2009) Generation of transgenic non-human primates with germline transmission. *Nature* 459:523–527
  49. McGrew MJ, Sherman AS, Ellard FM, Lillico SG, Gilhooley HJ, Mitrophanous KA, Kingsman AJ, Sang H (2004) Efficient production of germline transgenic chickens using lentiviral vectors. *EMBO Rep* 5:728–733
  50. Kim JN, Park TS, Park SH, Park KJ, Kim TM, Lee SK, Lim JM, Han JY (2010) Migration and proliferation of intact and genetically modified primordial germ cells and the generation of a transgenic chicken. *Biol Reprod* 82:257–262
  51. McGrew MJ, Lillico SG, Sherman A, Taylor L, Sang H (2010) Functional conservation between rodents and chicken of regulatory sequences driving skeletal muscle gene expression in transgenic chickens. *BMC Dev Biol* 10:e26
  52. Scott BB, Lois C (2005) Generation of tissue-specific transgenic birds with lentiviral vectors. *Proc Natl Acad Sci USA* 102:16443–16447
  53. Poynter G, Lansford R (2008) Generating transgenic quail using lentiviruses. *Methods Cell Biol* 87:281–293
  54. Shin SS, Kim TM, Kim SY, Kim TW, Seo HW, Lee SK, Kwon SC, Lee GS, Kim H, Lim JM, Han JY (2008) Generation of transgenic quail through germ cell-mediated germline transmission. *FASEB J* 22:2435–2444
  55. Agate RJ, Scott BB, Haripal B, Lois C, Nottebohm F (2009) Transgenic songbirds offer an opportunity to develop a genetic model for vocal learning. *Proc Natl Acad Sci USA* 106:17963–17967
  56. Kwon SC, Choi JW, Jang HJ, Shin SS, Lee SK, Park TS, Choi IY, Lee GS, Song G, Han JY (2010) Production of biofunctional recombinant human interleukin 1 receptor antagonist (rhIL1RN) from transgenic quail egg white. *Biol Reprod* 82:1057–1064
  57. McClintock B (1953) Induction of instability at selected loci in maize. *Genetics* 38:579–599
  58. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41:331–368



59. Ivics Z, Li MA, Mátés L, Boeke JD, Nagy A, Bradley A, Zsuzsanna I (2009) Transposon-mediated genome manipulation in vertebrates. *Nat Meth* 6:415–422
60. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T (2005) Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* 122:473–483
61. Balciunas D, Wangenstein KJ, Wilber A, Bell J, Geurts A, Sivasubbu S, Wang X, Hackett PB, Largaespada DA, McIvor RS, Ekker SC (2006) Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS Genet* 2:e169
62. Sherman A, Dawson A, Mather C, Gilhooley H, Li Y, Mitchell R, Finnegan D, Sang H (1998) Transposition of the *Drosophila* element mariner into the chicken germ line. *Nat Biotechnol* 16:1050–1053
63. Sato Y, Kasai T, Nakagawa S, Tanabe K, Watanabe T, Kawakami K, Takahashi Y (2007) Stable integration and conditional expression of electroporated transgenes in chicken embryos. *Dev Biol* 305:616–624
64. Lu Y, Lin C, Wang X (2009) PiggyBac transgenic strategies in the developing chicken spinal cord. *Nucleic Acids Res* 37:e141
65. Petitte JN, Clark ME, Liu G, Verrinder-Gibbins AM, Etches RJ (1990) Production of somatic and germline chimeras in the chicken by transfer of early blastodermal cells. *Development* 108:185–189
66. Etches RJ, Clark ME, Toner A, Liu G, Gibbins AM (1996) Contributions to somatic and germline lineages of chicken blastodermal cells maintained in culture. *Mol Reprod Dev* 45:291–298
67. Pain B, Clark M, Nakazawa H, Sakurai M, Samarut J, Etches R (1996) Long-term in vitro culture and characterization of avian embryonic stem cells with multiple morphogenetic potentialities. *Development* 122:2339–2348
68. Bezzubova O, Silbergleit A, Yamaguchi-Iwai Y, Takeda S, Buerstedde JM (1997) X-ray resistance and homologous recombination frequencies in a RAD54<sup>-/-</sup> mutant of the chicken DT40 cell line. *Cell* 89:185–193
69. van de Lavoie MC, Mather-Love C, Leighton P, Diamond JH, Heyer BS, Roberts R, Zhu L, Winters-Digiaccio P, Kerchner A, Gessaro T, Swanberg S, Delany ME, Etches RJ (2005) High-grade transgenic somatic chimeras from chicken embryonic stem cells. *Mech Dev* 123:31–41
70. Laval F, Acloque H, Bachelard E, Nieto MA, Samarut J, Pain B (2009) Ectopic expression of *Cvh* (Chicken *Vasa* homologue) mediates the reprogramming of chicken embryonic stem cells to a germ cell fate. *Dev Biol* 330:73–82
71. Tsunekawa N, Naito M, Sakai Y, Nishida T, Noce T (2000) Isolation of chicken *vasa* homolog gene and tracing the origin of primordial germ cells. *Development* 127:2741–2750
72. Eyal-Giladi H, Ginsburg M, Farbarov A (1981) Avian primordial germ cells are of epiblastic origin. *J Embryol Exp Morphol* 65:139–147
73. Simkiss K, Rowlett K, Bumstead N, Freeman BM (1989) Transfer of primordial germ cell DNA between embryos. *Protoplasma* 151:164–166
74. Chang IK, Jeong DK, Hong YH, Park TS, Moon YK, Ohno T, Han JY (1997) Production of germline chimeric chickens by transfer of cultured primordial germ cells. *Cell Biol Int* 21:495–499
75. Vick L, Li Y, Simkiss K (1993) Transgenic birds from transformed primordial germ cells. *Proc Biol Sci* 251:179–182
76. Mozdziaik PE, Wysocki R, Angerman-Stewart J, Pardue SL, Petitte JN (2006) Production of chick germline chimeras from fluorescence-activated cell-sorted gonocytes. *Poult Sci* 85:1764–1768
77. Tajima A, Naito M, Yasuda Y, Kuwana T (1993) Production of germ line chimera by transfer of primordial germ cells in the domestic chicken (*Gallus domesticus*). *Theriogenology* 40:509–519
78. Ono T, Machida Y (1999) Immunomagnetic purification of viable primordial germ cells of Japanese quail (*Coturnix japonica*). *Comp Biochem Physiol A Mol Integr Physiol* 122:255–259
79. Zhao DF, Kuwana T (2003) Purification of avian circulating primordial germ cells by nycodenz density gradient centrifugation. *Br Poult Sci* 44:30–35
80. Park TS, Jeong DK, Kim JN, Song GW, Hong YH, Lim JM, Han JY (2003) Improved germline transmission in chicken chimeras produced by transplantation of gonadal primordial germ cells into recipient embryos. *Biol Reprod* 68:1657–1662
81. Kalina J, Senigl F, Micáková A, Mucksová J, Blazková J, Yan H, Poplstein M, Hejnar J, Trefil P (2007) Retrovirus-mediated in vitro gene transfer into chicken male germ line cells. *Reproduction* 134:445–453
82. Park TS, Hong YH, Kwon SC, Lim JM, Han JY (2003) Birth of germline chimeras by transfer of chicken embryonic germ (EG) cells into recipient embryos. *Mol Reprod Dev* 65:389–395
83. van de Lavoie MC, Diamond JH, Leighton PA, Mather-Love C, Heyer BS, Bradshaw R, Kerchner A, Hooi LT, Gessaro TM, Swanberg SE, Delany ME, Etches RJ (2006) Germline transmission of genetically modified primordial germ cells. *Nature* 441:766–769
84. Leighton PA, van de Lavoie MC, Diamond JH, Xia C, Etches RJ (2008) Genetic modification of primordial germ cells by gene trapping, gene targeting, and  $\phi$ C31 integrase. *Mol Reprod Dev* 75:1163–1175
85. Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, Hocking TD, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Amacher SL (2008) Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat Biotechnol* 26:702–708
86. Geurts AM, Cost GJ, Rémy S, Cui X, Tesson L, Usal C, Ménoret S, Jacob HJ, Anegón I, Buelow R (2010) Generation of gene-specific mutated rats using zinc-finger nucleases. *Methods Mol Biol* 597:211–225
87. Petitte JN (2006) Avian germplasm preservation: embryonic stem cells or primordial germ cells? *Poult Sci* 85:237–242
88. Muñoz M, Trigal B, Molina I, Díez C, Caamaño JN, Gómez E (2009) Constraints to progress in embryonic stem cells from domestic species. *Stem Cell Rev* 5:6–9

## Books and Reviews

- Nieuwkoop PD, Sutasurya LA (1979) Primordial germ cells in the chordates. Cambridge University Press, Cambridge
- Petitte JD (2004) Isolation and maintenance of avian ES cells. In: Handbook of stem cells, vol 1, Embryonic stem cells. Elsevier, Amsterdam, pp 471–479, Chapter 44
- Pfeifer A, Hofmann A (2009) Lentiviral transgenesis. *Methods Mol Biol* 530:391–405
- Walsh G (2003) Biopharmaceuticals: biochemistry and biotechnology, 2nd edn. Wiley, Chichester

## Aviation and Atmosphere

HARTMUT GRASSL

Max Planck Institute for Meteorology, Hamburg,  
Germany

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Emissions of Climatically Relevant Substances by  
 Aviation  
 Aviation-Induced Cloudiness  
 Changes of Air Chemistry by Air Traffic  
 Transport Sector's Contribution to Future Climate  
 Change  
 Why Did International Air Traffic So Far Escape from  
 an Emission Reduction Policy?  
 The European Directive to Include Air Traffic into the  
 Emission Trading Scheme  
 An Airline Emission Index  
 Conclusions and Outlook  
 Acknowledgment  
 Bibliography

### Glossary

- AIC** Aviation-induced cloudiness is the sum of all changes in cloudiness associated with aviation operations hence, it combines contrails and so-called persistent contrail cover.
- Airline emission index** A measure to compare the performance of different airlines with respect to emissions relative to an ideal one.

**Aviation** Here: commercial air traffic.

**Contrails** Jet engine exhaust condensation trails of airplanes.

**Emission trading** A market-based approach used to control pollution by providing economic incentives for achieving reductions in the emissions of pollutants.

**GWP** The global warming potential is a relative measure of how much heat a greenhouse gas traps in the atmosphere. It compares the amount of heat trapped by a certain mass of the gas in question to the amount of heat trapped by the same mass of carbon dioxide.

**GTP** The GTP is defined as the ratio between the global mean surface temperature change at a given future time horizon following an emission (pulse or sustained) of a compound relative to a reference gas.

**Radiative forcing** When the balance between the radiative energy received from the sun and radiative energy leaving the Earth-atmosphere system is perturbed, e.g., by a concentration change of an absorbing gas, the resulting imbalance is termed radiative forcing that initiates climate change.

### Definition of the Subject

Commercial air traffic's peculiarity, emissions into the atmosphere mainly at cruising altitude, leads to changed impacts on global climate in comparison to other modes of transport. This entry concentrates on these differences but also looks onto measures to reduce emissions, e.g., the emission trading scheme in the European Union, and presents an airline emission index comparing commercial airlines.

### Introduction

The exhaust of commercial and also military aircraft needs – in comparison to the one from power plants, cars, trucks, heaters in buildings, railway, ships, and burners in industry – a special discussion, because of its main injection at cruising altitudes, i.e., into the upper troposphere or lower stratosphere from about 9 to 12 km height. This part of the Earth's atmosphere is nearly all the time colder than  $-40^{\circ}\text{C}$ , often it is with about  $-70^{\circ}\text{C}$  the coldest one in mid-latitudes in the lowest at least 60 km. Therefore, the water vapor

concentration in these layers must be lower by typically 3–4 orders of magnitude than at the Earth's surface in low- and mid-latitudes. Hence, several emissions of an aircraft besides carbon dioxide, for example unavoidable water vapor, in principle avoidable nitrogen oxides can lead to a considerable disturbance of the radiation budget of the planet. First, via the formation of rather persistent condensation trails, an ice cloud, during frequently favorable meteorological conditions, and second, via the formation of ozone around cruising altitudes because of the nitrogen oxide emissions.

In addition, the growth rate of commercial air traffic was on average above the one for the global gross economic product, leading to a growing relative contribution of air traffic to the changed composition of the atmosphere. Only very recently did the growth rate no longer surmount the growth rate of world economy as strong as before.

However, it was only in the late 1980s when first publications dealt with air traffic's potential influence on *climate*. Earlier, in the early 1970s, the planned supersonic aircraft fleet in the USA had led to an assessment of the impact of such aircraft on the stratospheric ozone layer. At this time the consequences of the emission of the trace gas nitrogen dioxide (NO<sub>2</sub>) were the main reasons for the investigations, because it is strongly involved in stratospheric ozone chemistry leading to ozone depletion at altitudes above about 20 km but to ozone increases in the troposphere and lowest stratosphere.

In 1996, the steadily growing public and scientific discussions on air traffic and its potential effect on global climate have stimulated the International Civil Aviation Organization (ICAO), the UN specialized agency that has global responsibility for the establishment of standards, recommended practices, and guidance on various aspects of international civil aviation, including environmental protection, to ask the Intergovernmental Panel on Climate Change (IPCC) to assess in a special report [17] entitled "Aviation and the Global Atmosphere" the air traffic influence on global climate. In this report not only the effects of air traffic on atmospheric composition and on climate were assessed but also projected into the future using air traffic growth scenarios. These results and new findings have entered the third and fourth full IPCC assessments [19–21]; hence, the hitherto known facts have been presented to

the political arena and have led to first hesitating measures to reduce emissions, like the European Union's decision in 2009 to include air traffic into the European emissions trading scheme, to become effective on January 1, 2012.

This entry will start with the greenhouse gas emission history of air traffic in section "[Emissions of Climatically Relevant Substances by Aviation](#)," influenced both by the above average growth rate of this economic sector but also by the strong increase in fuel use efficiency per passenger kilometer, which has been stronger than for road traffic and much stronger than for railways. This section will also include projections of future air traffic greenhouse gas emissions using estimates from scenarios of IPCC. It continues with the knowledge about long-lasting contrails and their dependence on favorable meteorological conditions in section "[Aviation-Induced Cloudiness](#)." This section will also include cirrus clouds formed or transformed because of air traffic, existing besides the easily detectable linear contrails. Section "[Changes of Air Chemistry by Air Traffic](#)" is then devoted to changes in air chemistry caused by air traffic emissions. Section "[Transport Sector's Contribution to Future Climate Change](#)" extrapolates into the 21st century with known scenarios of human behavior. It follows the section on "[Why did International Air Traffic So Long Escape an Emissions Trading Scheme?](#)" The first attempt of the European Union to include international air traffic into an emissions trading scheme will be described in section "[The European Directive to Include Air Traffic into the Emission Trading Scheme](#)." The second last section on "[An Airline Emission Index](#)" will present essentials of an airline index that includes a ranking concerning carbon dioxide and nitrogen oxides emissions, published very recently. The last section will provide the conclusions and also present avenues how to reduce or even to avoid air traffic emissions changing climate.

### **Emissions of Climatically Relevant Substances by Aviation**

The estimates of climate effects of air traffic depend totally on a reliable time-dependent emission inventory on global scale for the past, the present, and the future. The major research project QUANTIFY of

the European Union's Sixth Framework Research Programme (finished very recently) has established such a consistent global three-dimensional air transport emission inventory for greenhouse gases, particles, and precursor gases of other greenhouse gases and for aerosol particles for the year 2000 (also for all other modes of transport), and has developed estimates of future emissions for the years 2020, 2050, and 2100. The example for air traffic is shown in Fig. 1 and Table 1, taken from [25].

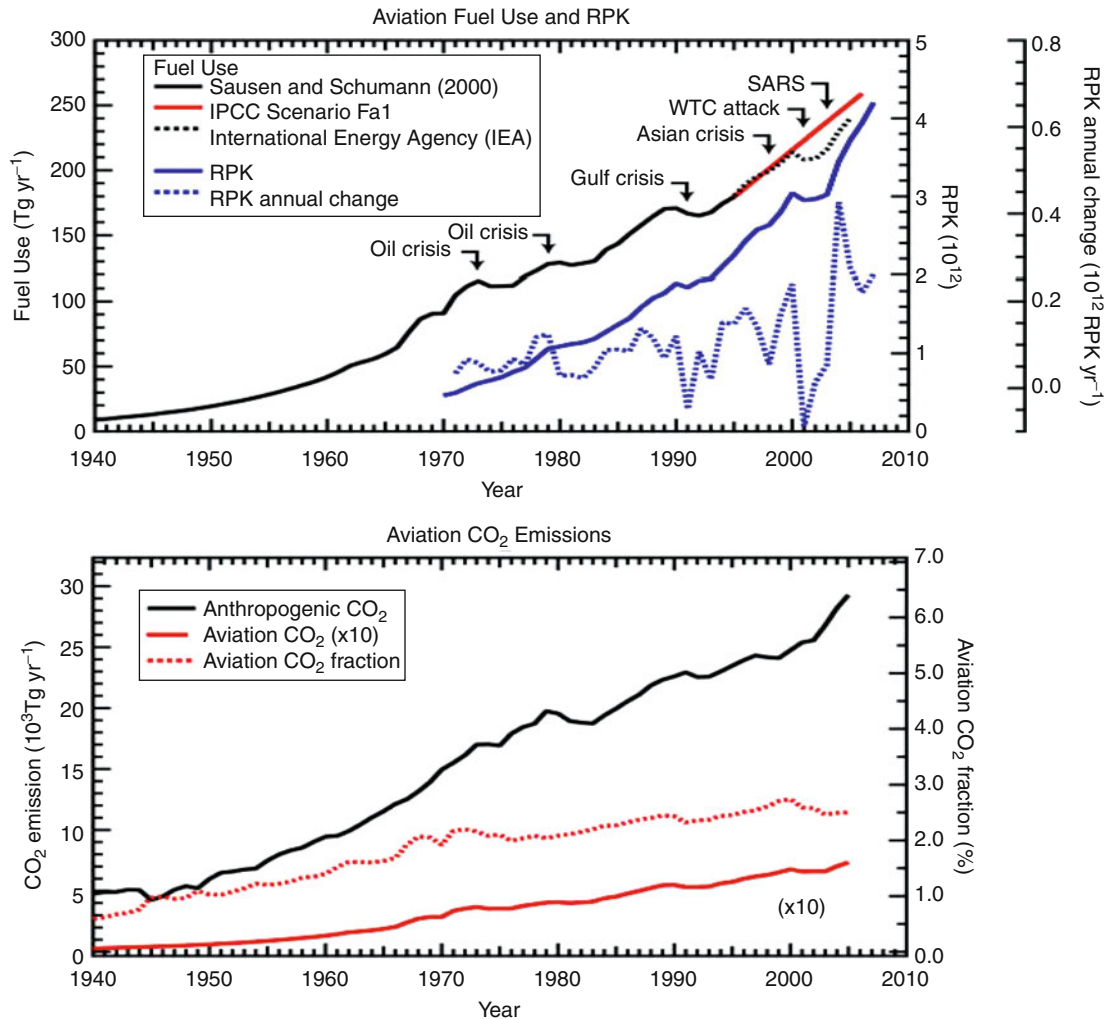
Since different substances in the aircraft exhaust have very different concentrations and also very different radiative properties, a common scale must be created for a just assessment of the environmental burden caused by different aircraft and flight types. The next subsection will deal with it.

### Creating a Common Scale for the Climate Effects of Air Traffic Emissions

The entire motorized transport sector (road traffic, shipping, rail, aviation) emits a wide variety of gases and aerosols into the atmosphere, with different characteristics, which either influence climate directly by their radiative properties or indirectly via chemical and physical processes influencing concentrations of other climatically relevant substances. We need metrics that allow these emissions to be placed on a common scale in terms of their impact on global climate. The applications of such metrics are manifold: First, they have served and will serve as basic information in negotiations for international agreements and emission trading schemes (see section "The European Directive to Include Air Traffic into the Emission Trading Scheme"); second, they allow the assessment of trade-offs between changes in emissions resulting from technological or operational developments; third, they serve for comparisons of different environmental impacts caused by the different transport sectors and thus help in traffic policy making. Usually, the CO<sub>2</sub> emission, common to all transport sectors, is used as a reference. Therefore, all other emissions have to be transformed into their "CO<sub>2</sub>-equivalents." Ideally, the same equivalent CO<sub>2</sub> emissions should produce the same climate effect, regardless of their shares in the exhaust or after chemical transformation in the atmosphere. This is, however, only an approximation of

reality, because most of the non-CO<sub>2</sub> emissions from the different transport sectors, and especially aviation, are short-lived substances with often rather regionalized concentrations. In addition, these short-lived substances are all not included in the Kyoto Protocol basket of solely long-lived greenhouse gases.

The challenges in developing generally accepted metrics are strongly magnified by the inclusion of short-lived species. One first difficulty concerns the choice of an appropriate structure for the metric. Should effects of an emission today be integrated over 10, 20, 50, 100, or even 200 years? Or should the impact, e.g., the global mean surface temperature increase after a certain time period, be used as a measure? The first approach is the most frequently used one, characterized by the global warming potential (GWP) and the one adopted in the Kyoto Protocol. The second one, called global temperature change potential (GTP), strongly reduces the influence of short-lived emissions if the same long time horizon is used as typically for GWP. In view of the lifetime (defined as the time for the decay of the concentration of an emitted pulse of a substance to 1/e) in the atmosphere of an anthropogenic addition of CO<sub>2</sub>, which is well above 100 years, 100 years have been chosen for the Kyoto Protocol, although there was no specific argument put forward for this time horizon by IPCC. A further challenge is the quantification of input parameters for the metric, originating in the uncertainty of atmospheric processes. For example: How much of the nitrogen oxide emission remains available for the formation of ozone in the troposphere, a potent but short-lived greenhouse gas of the atmosphere? On top of these challenges we have to face the dependence of the effects on the location of the emissions for all short-lived substances, severest for air traffic, where for example the emission height is decisive for the formation of a long-lasting contrail. The QUANTIFY project has evaluated existing metrics and has introduced new concepts (for details see Fuglestvedt et al. 2009). All these metrics have been evaluated for emissions from the transport sector both in terms of integrated radiative forcing (When the balance between the radiative energy received from the sun and radiative energy leaving the Earth-atmosphere system is perturbed, e.g., by a concentration change of an absorbing gas, the



**Aviation and Atmosphere. Figure 1**

*Upper Panel:* Aviation fuel usage beginning in 1940 from Sausen and Schumann [32] and extended with data from IEA [16] and the IPCC Fa1 scenario of Henderson et al. [13]. The arrows indicate world events that shortly threatened global aviation use: the oil crises of the 1970s, the Gulf war crisis in the early 1990s, the Asian financial crisis in the late 1990s, the World Trade Center (WTO) attack in 2001, and the global health crisis brought about by the severe acute respiratory syndrome (SARS). Also shown is the growth in air passenger traffic from 1970 to 2007 in trillions of revenue passenger kilometers (RPK) (near right-hand axis) (Source: ICAO traffic statistics from <http://www.airlines.org/economics/traffic/World/Airline/Traffic.htm>. Accessed 19 Sept 2007) and the annual change in RPK (far right-hand axis [Note offset zero]).

*Lower Panel:* Growth in CO<sub>2</sub> emissions in  $\text{Tg CO}_2/\text{year}$  for all anthropogenic activities and from aviation fuel burn (left-hand axis), and the fraction of total anthropogenic CO<sub>2</sub> emissions represented by aviation CO<sub>2</sub> emissions (%) (right-hand axis). Note scaling of aviation CO<sub>2</sub> emissions (Source: Lee et al. [24])

resulting imbalance is termed radiative forcing. It can be calculated by changing the concentration of the substance and keeping all other parameters fixed. In the real climate system, the imbalance stimulates

climate change trying to restore the energy balance of the planet.) (GWP) and the net surface temperature change (GTP) of the various transport modes for several time horizons. For the series of tasks to be



**Aviation and Atmosphere. Table 1** Estimations of historical and current emissions from inventory models, 1990–2005

| Year | Inventory | Civil fuel Tg/a        | Military fuel Tg/a | Total fuel Tg/a | Total CO <sub>2</sub> Tg/a | Distance (civil, 10 <sup>9</sup> km) | NO <sub>x</sub> TgNO <sub>2</sub> /a | EINO <sub>x</sub> g NO <sub>x</sub> /kg fuel |
|------|-----------|------------------------|--------------------|-----------------|----------------------------|--------------------------------------|--------------------------------------|----------------------------------------------|
| 1990 | FAST      | 105                    | n.c.               |                 | 331                        | 15.7                                 | 1.42                                 |                                              |
| 1992 | NASA      | 114                    | 25.5               | 139             | 359                        | n.c.                                 | 1.44                                 | 12.6(12.0)                                   |
| 1992 | ANCAT/EC2 | 114                    | 17.1               | 131             | 360                        | n.c.                                 | 1.60(1.81)                           | 14.0(13.8)                                   |
| 1999 | NASA      | 128 (134 inc. charter) | n.c.               |                 | 404 (423 inc. charter)     | 25.8                                 | 1.69                                 | 13.2                                         |
| 2000 | TRADEOFF  | 152                    | 17.1 <sup>a</sup>  | 169             | 480                        | 25.1                                 | 1.95(2.15)                           | 12.7(12.9)                                   |
| 2000 | FAST      | 152                    | n.c.               |                 | 480                        | 26.9                                 | 2.03                                 | 13.4                                         |
| 2000 | SAGE      | 181                    | n.c.               |                 | 572                        | 33.3                                 | 2.51                                 | 13.8                                         |
| 2001 | SAGE      | 170                    | n.c.               |                 | 536                        | 31.9                                 | 2.35                                 | 13.8                                         |
| 2002 | SAGE      | 171                    | n.c.               |                 | 539                        | 32.6                                 | 2.41                                 | 14.1                                         |
| 2002 | AERO2K    | 156                    | 19.5               | 176             | 492(553)                   | 33.2                                 | 2.06(2.24)                           | 13.2(12.7)                                   |
| 2003 | SAGE      | 176                    | n.c.               |                 | 557                        | 34.5                                 | 2.49                                 | 14.1                                         |
| 2004 | SAGE      | 188                    | n.c.               |                 | 594                        | 37.0                                 | 2.69                                 | 14.3                                         |
| 2005 | SAGE      | 203                    | n.c.               |                 | 641                        | 40.7                                 | 2.90                                 | 14.2                                         |

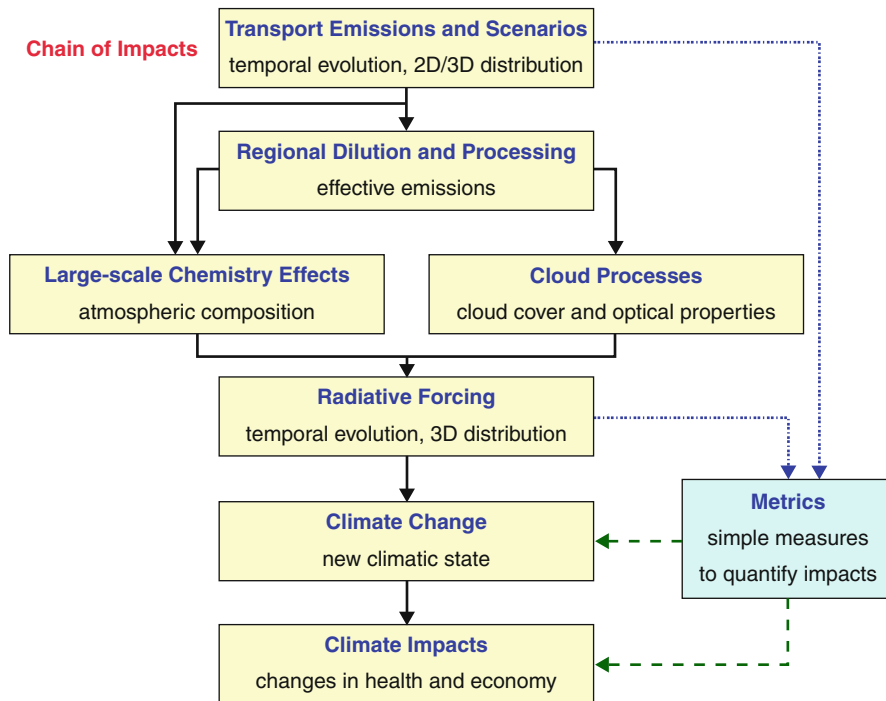
References: "FAST," Lee et al. (2005); "NASA," (Baughcum et al. [3]) and (Sutkus et al. [33]); "ANCAT/EC2," Gardner et al. [12]; "TRADEOFF," Gauss et al. (2006); "SAGE," Kim et al. [22]; "AERO2K," Eyers et al. [10]

<sup>a</sup>Military emissions adopted from other study; n.c. not calculated, brackets include military emissions

performed starting from transport sector emissions and ending at climate impacts, see Fig. 2 [30].

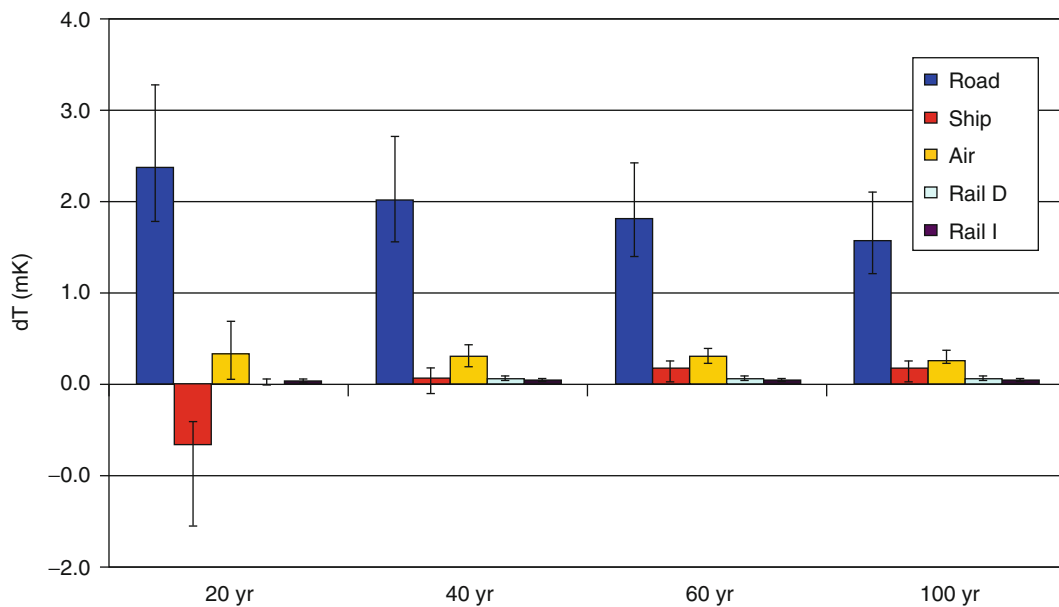
For metrics, very different approaches are possible. One can look at the current radiative or temperature forcing relative to the preindustrial era or at the temporal development of the forcing over several decades for a pulse emitted today, or to the forcing for a sustained emission or a more elaborate emission scenario. Figure 3 shows the temperature change that will be caused by the emissions of the year 2000 for all transport modes for four chosen time horizons. In order to calculate surface air temperature changes ideally a coupled Atmosphere/Ocean/Land-Model has to be run. If this model gets as an input the three-dimensional emission distribution of air traffic, it must not lead to the same relation of different transport sectors as the radiative forcing calculation, because it adds the processes at the ocean air interface that are neglected in the chemical transport model used for the radiative transfer calculations because then

sea surface temperature is fixed. Figure 4 clearly demonstrates that emissions from road transport in 2000 are dominating the traffic sector impact on surface temperature, with aviation as the second largest contributor to surface air temperature changes. For time horizons of 20 and 100 years, the net warming from road transport is seven and six times larger than the net warming from aviation. This is approximately the same ratio as the ratio between the CO<sub>2</sub> emissions from these sectors; thus, mainly CO<sub>2</sub> controls the climate response on timescales beyond 10 years for the GTP metric. The net temperature effect of shipping changes sign from negative for a time horizon of 20 years to positive for the longer time horizons. Both rail sectors' contributions seem small, but added together they reach 15–30% of the effect of aviation for time horizons between 20 and 100 years. The results in Fig. 3 show well how the relative importance of the different transport sectors changes depending on the time perspective.



**Aviation and Atmosphere. Figure 2**

From emissions to climate change and climate impacts (a schematic). Only the main fluxes of information are indicated by arrows. QUANTIFY science tasks are printed in blue, output in black (Source: QUANTIFY [30])



**Aviation and Atmosphere. Figure 3**

Contribution from a 1-year pulse of current (year 2000) emissions to net future temperature change (mK) for each transport mode for four time horizons (20, 40, 60, and 100 years), including the uncertainties, here given by the 1-sigma level (Source: Lee et al. [25])

## Climate Change Contribution of Carbon Dioxide Emissions from Air Traffic

The entire transport sector contributes about one third of the total global anthropogenic CO<sub>2</sub> emissions. The annual growth rate of transport-related greenhouse gas emissions is larger than for other mature industrial sectors. In view of the United Nations Framework Convention on Climate Change (UNFCCC), its Kyoto Protocol, and possible follow-up protocols, this high rate of increase creates the severest problem for the aviation sector, when trying to achieve forthcoming emission reduction targets. Additionally, the impact of the transport sector on climate is at least as complex as for other sectors like heating of houses, because it also involves more strongly air chemistry and modification of cloudiness than the emissions from shipping, and in addition direct cloud formation (contrails) by air traffic. Thus an assessment of the full climate effects of the transport sector and especially aviation has to take into account more than just the list of gases in the Kyoto Protocol. The transport sector can impact on climate in the following ways:

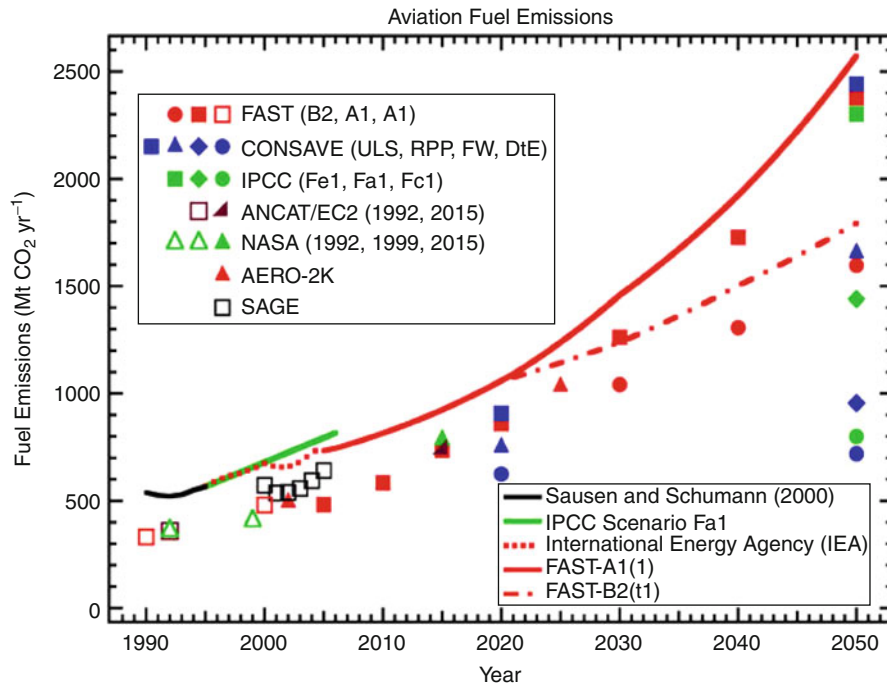
- By direct emissions of greenhouse gases, mainly CO<sub>2</sub>, but also by nitrous oxide (N<sub>2</sub>O) and some halocarbons
- By emissions of precursors of ozone, such as nitrogen oxides (NO<sub>x</sub>), carbon monoxide (CO), and volatile organic compounds (VOCs)
- By emissions of aerosol particles or their precursor gases, in particular black carbon (soot) and sulfur dioxide (SO<sub>2</sub>), which are directly and indirectly (via cloud formation and cloud modification) radiatively and chemically active
- By directly triggering additional clouds (e.g., contrails and contrail cirrus)

On the other hand, the global and economic system is largely dependent on an efficient transport system. This dependency has grown during recent decades not only because of an increasing world population but mainly because of globalization. Life style and the availability of cheaper transport (mainly because of lack of internalization of external costs) in developed and developing countries are also enhancing demand. Further very strong growth of the transport sector is expected in developing countries. In the long term, a

sustainable transport system is needed that satisfies the demands of economy and population while following the constraints of climate change policy measures. In order to meet these constraints clear information on the climatic impact of different transport emissions is needed. Here we concentrate on aviation, a smaller part with respect to goods and passengers transported, but with many climatically relevant facets because of its already mentioned main emission height of about 10 km.

Each kilogram of aviation fuel, overwhelmingly kerosene, when burned in an engine, must lead to 3.16 kg carbon dioxide (CO<sub>2</sub>) and 1.0 kg of water vapor (H<sub>2</sub>O). These unavoidable emissions are two major reasons of an air traffic influence on the radiative transfer in the atmosphere, but not the only ones. The first one is nearly completely independent of the place where the emission occurs, because of the long lifetime of anthropogenic carbon dioxide being well above 100 years and thus resulting in a globally well-mixed anthropogenic carbon dioxide concentration addition. Aviation CO<sub>2</sub> will contribute to the entire CO<sub>2</sub> burden of the atmosphere like all other emissions into the atmosphere close to the surface of the Earth. For the second major emission, H<sub>2</sub>O, the point where it is emitted is decisive: At the surface this addition is negligible in comparison to the natural fluxes of water, but not so at about 10 km height at a water vapor mixing ratio ranging from about 100 ppmv to as low as 5 ppmv. Flying by 1 km higher or lower in heights around the tropopause (the interface between the troposphere and the stratosphere; typically at 10 km height in mid-latitudes), however, can often avoid a long-lasting linear contrail (for more see section “[Aviation Induced Cloudiness](#)”).

At present the global consumption of aviation fuel per year is equivalent to an emission of about 0.2 Gigatonnes of Carbon per year (GtC/a), which is about 2% of the entire anthropogenic carbon emissions, when emissions from land use change (deforestation being the main contribution to it) are included. If only counting the emissions originating from the burning of fossil fuels and cement production, the percentage of aviation emissions grows to about 2.4. [Figure 4](#) presents an estimate of the time series of global CO<sub>2</sub> emissions by aviation since 1990 and extrapolation attempts by different groups as compiled



**Aviation and Atmosphere. Figure 4**

Historical and present-day inventories, and future projections of civil aviation CO<sub>2</sub> emissions from a variety of sources: AERO2K [10]; ANCAT/EC2 [12]; CONSAVE [5]; FAST (Owen and Lee 2006); IPCC [17]; NASA [3, 4, 33]; SAGE [22]. The *open symbols* indicate inventory analysis and the *closed symbols* indicate projections. Also shown are the CO<sub>2</sub> emissions implied by IEA fuel sales statistics [16]. The IEA data represent the total of civil and military usage because all kerosene sales are included. The Sausen and Schumann [32] data are also based on IEA. The solid (*dashed*) lines for FAST-A1 (B2) scenarios (evaluated with the t1 technology option) and the IPCC Fa1 scenario also account for all fuel sales in order to be consistent with the IEA values ending in 2005. In the figure legend, the FAST, CONSAVE, and IPCC symbols are shown in an order that matches the scenario labels in the parentheses in each case (Source: Lee et al. [24])

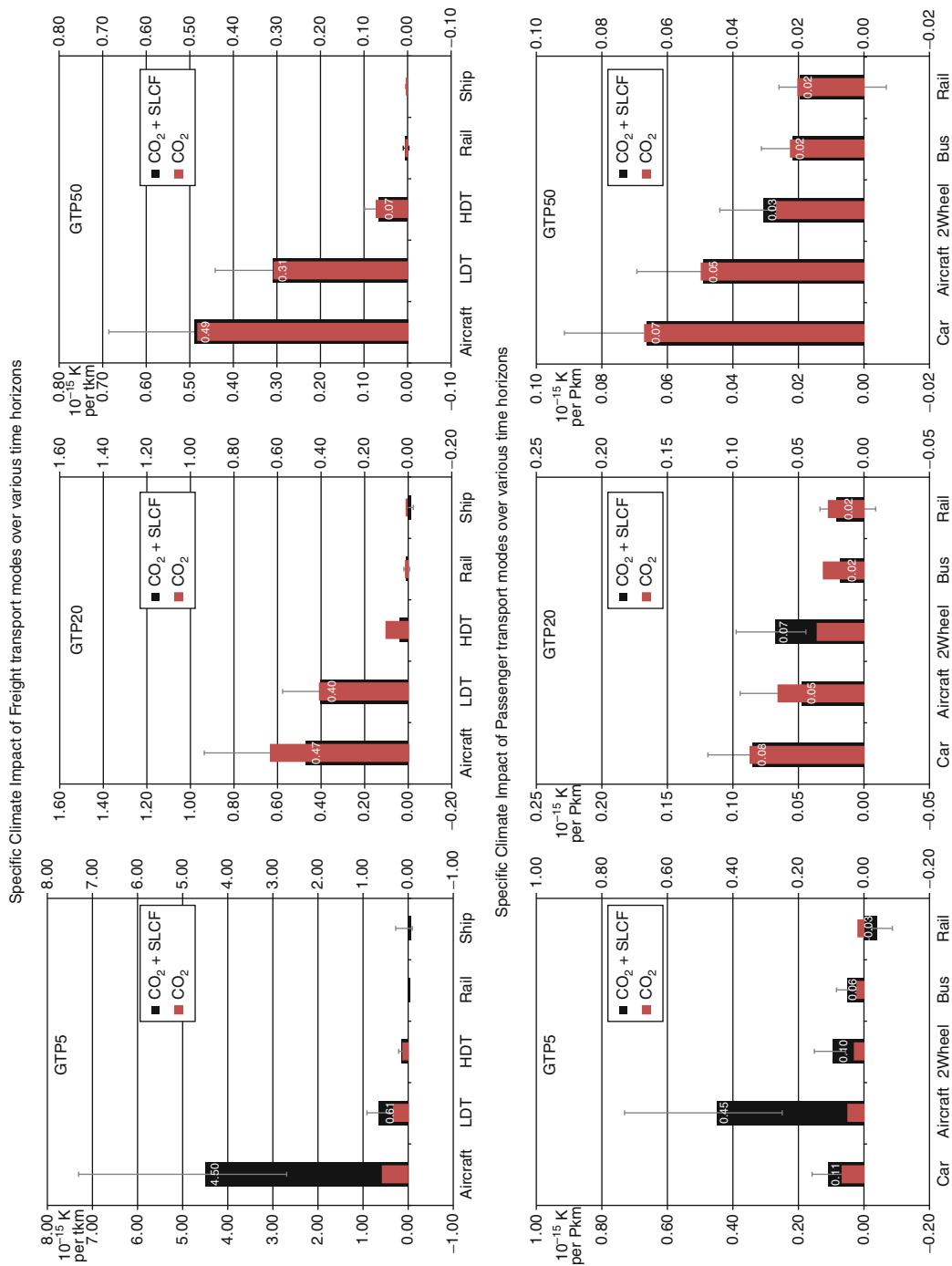
by Lee et al. [25]. Besides an overall growth of the emissions, it is evident that different scenarios of future economic development and technological progress lead to very different emission paths.

It is also interesting to compare the different modes of transport with respect to emissions per passenger and per tonne transported. Here again results from the recently finished major European Union's research project QUANTIFY [30], to which the author has contributed, are presented: Aviation is the strongest contributor to radiative forcing by CO<sub>2</sub> emissions per passenger kilometer in comparison to road traffic, shipping, and rail. As clearly visible from Fig. 6, this is not the case for freight transport, where road

traffic contributes similarly to radiative forcing per tonne-kilometer (Fig. 5).

### Aviation-Induced Cloudiness

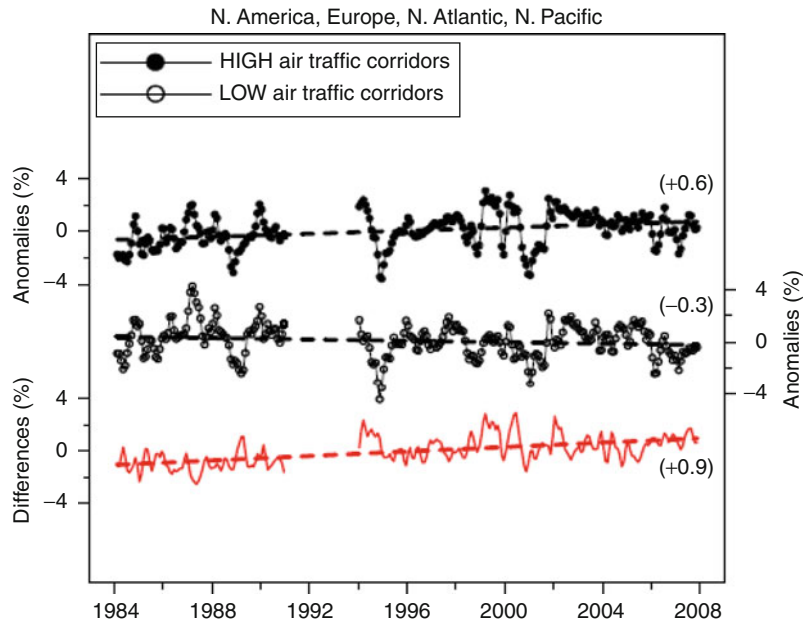
There exist two types of impact on clouds caused by particle emissions from traffic. First, the particle emissions can cause the formation of clouds that would not form naturally. The most prominent example is jet exhaust condensation trails (contrails). This cloud circumvents the lack of so-called freezing nuclei in the rather clean and cold upper troposphere. The large emission of water vapor (1 kg/l of kerosene burned) in the hot exhaust plume together with the additional particles, also within the exhaust



Aviation and Atmosphere. Figure 5

Temperature change caused by emissions in the year 2000 after 5, 20, and 50 years; For passenger transport modes (*upper panels*) per passenger kilometer, for freight transport modes (*lower panels*) per tonne-kilometer. The columns differentiate the impact when all short- and long-lived climate forcings (in *black*) are summed up and when only CO<sub>2</sub> (in *color*) is considered (Source: QUANTIFY [30])





**Aviation and Atmosphere. Figure 6**

Averaged cirrus cloud trends over high and low air traffic regions in North America, Europe, North Atlantic, and North Pacific (Source: QUANTIFY [30])

plume, leads to cloud droplet formation, which rapidly freeze by mixing with the ambient air. If the flight corridor of an aircraft is supersaturated with respect to an ice surface, the contrail formed will survive and expand depositing all the supersaturated water vapor onto the contrail ice particles.

A second type of anthropogenic cirrus clouds, which is much more difficult to observe, are aerodynamic contrails involving a sequence of effects: The emitted particles first get mixed with the background aerosol particles and later – when the mix of anthropogenic and natural aerosol particles becomes involved into cloud formation processes – the particles emitted by air traffic will affect cloud formation, cloud evolution, cloud properties, and eventually precipitation. Also most so-called ship tracks belong to this category. A variant of this second aviation-induced cloud type, the hypothetical soot cirrus, is of interest here. It describes an ice cloud that is strongly affected by soot particles emitted by aircraft. To estimate the effect of aerosol particles emitted by air traffic (often soot) on ice clouds is difficult, because of the presence of various pathways to ice formation. In clean unpolluted air, ice

clouds form via homogeneous nucleation, while in polluted air, like in air traffic corridors, heterogeneous nucleation might be dominant. Because there exist always very many other aerosol particles, which become solution droplets at higher relative humidity over a water surface (about  $100/\text{cm}^3$ ), an ice cloud forming homogeneously has a high number concentration of ice crystals. In contrast, heterogeneous ice cloud formation is a very selective process: Only about every millionth particle is an appropriate ice forming nucleus (IN). Thus, ice clouds forming predominantly by heterogeneous nucleation have a small number concentration of ice crystals. IN injection into a clean environment leads to heterogeneous ice crystal formation *before* homogeneous freezing can occur. Thus the ice cloud changed by air traffic has less but bigger crystals than a cloud that would have formed purely homogeneously. If heterogeneous ice formation is dominant as in a polluted air mass, the addition of IN can lead to more but smaller ice crystals. This simplified picture has to be modified: First, whether heterogeneous or homogeneous freezing dominates depends not only on the relative number of the respective

aerosol particle types but also on vertical wind speed (determining adiabatic cooling rates). Second, IN from various sources may have a spectrum of threshold supersaturations over an ice surface. These thresholds depend on temperature and on particle coating. Third, ice formation in supersaturated air with respect to an ice surface is difficult to be represented in large-scale models, because a supersaturated state is not uniquely related with a cloudiness state. Hence first model simulations like those by Penner et al. [29] have therefore rather the value of highlighting the problems than yielding firm results.

But also natural cirrus clouds can be affected by contrails formed within them or in their vicinity, in particular when these contrails spread and eventually form the so-called contrail cirrus, because these two ice cloud types compete for the same space and for the same supersaturated water vapor.

Aviation-Induced Cloudiness (AIC) has been defined as the sum of all changes in cloudiness associated with aviation operations [20]; hence, it combines jet exhaust condensation trails, often abbreviated as contrails, and so-called persistent contrail cover. The assessment of the latter is very difficult,

because AIC may be indistinguishable from background cirrus cover. This was the reason for IPCC [20] not to give a best estimate of AIC amounts and its associated radiative forcing. The value given in IPCC [20] for emissions until 2005 and a time horizon of 100 years, namely 0.03 W/m, is above the radiative forcing of 0.025 W/m for CO<sub>2</sub> alone. Thus more than a factor of 2 would be needed, if the total air traffic influence on radiative forcing has to be accounted for, knowing only the fuel burned and excluding additions related to air chemistry. Have these values given by IPCC in 2007 changed during recent research? Yes, but earlier relations between the impacts by greenhouse gases and cloudiness changes (caused by contrails and contrail cirrus) have not been altered significantly [25]. There is, however, a discussion about a new potential influence of aviation on cloudiness, the so-called soot cirrus (as already discussed above), relating to effects of changed optical properties of ice clouds into which soot from aircraft exhaust is mixed and which may act after transformations in the atmosphere as ice nuclei. The related model estimates [26, 29] diverge so strongly (see also the values given for soot cirrus in Table 2) and they are not backed by observations that

**Aviation and Atmosphere. Table 2** Year 2000 radiative forcing (in mWm<sup>-2</sup>) for each transport sector and each forcing mechanism relative to pre-industrial times, together with uncertainties. SWV is stratospheric water vapor

|                          | Road     | Ship      | Air          | Rail       |
|--------------------------|----------|-----------|--------------|------------|
| CO <sub>2</sub>          | 131 ± 15 | 34 ± 6    | 20 ± 3       | 23 ± 7     |
| Ozone                    | 32 ± 9   | 24 ± 6    | 18 ± 3       | 2 ± 0.4    |
| CH <sub>4</sub> lifetime | -7 ± 4   | -19 ± 2   | -5 ± 1       | -1 ± 0.4   |
| CH <sub>4</sub> – ozone  | -3 ± 2   | -8 ± 1    | -2 ± 0.5     | -0.3 ± 0.2 |
| SWV methane              | -1 ± 0.7 | -3 ± 0.7  | -1 ± 0.3     | -0.1 ± 0.1 |
| SWV direct               |          |           | 1 ± 0.2      |            |
| CFCs/HFCs                | 28 ± 3   |           |              |            |
| Stratospheric ozone      | -2 ± 4   |           |              |            |
| Sulphate aerosol         | -9 ± 4   | -26 ± 11  | -1 ± 0.4     | -1 ± 0.3   |
| Black carbon aerosol     | 44 ± 17  | 2 ± 1     | 0.3 ± 0.1    | 1 ± 0.3    |
| Organic carbon aerosol   | -2 ± 1   | -1 ± 0.5  |              | -0.5 ± 0.3 |
| Indirect aerosol effect  | 6 ± 111  | -49 ± 209 | -127 ± 128   | 0.5 ± 10   |
| Contrails                |          |           | 7 (-4, +6)   |            |
| Aviation induced cirrus  |          |           | 21 (-9, +17) |            |

much more research is needed before conclusions on a changed aviation policy could be drawn. In this context, it is too early to ask whether aviation's influence on radiative forcing might even change the sign.

Man-made changes in cirrus cloud cover can be detected by comparing cirrus trends over adjacent regions with dense and nearly no air traffic. Some exemplary long-term trends are shown in Fig. 6, taken from QUANTIFY [30]. It is evident that cirrus cloud coverage increases more or decreases less in regions with dense air traffic compared to regions with little air traffic. This could be considered a signature of increasing air traffic in the past 20 years. Such preliminary estimates from the differences in cirrus cloud cover, however, suggest an effect on radiative forcing of the order of only 6 mW/m for an assumed optical thickness of 0.4 for the cirrus clouds with air traffic influence.

Much better information on the statistics of optical thicknesses of single contrails, including the variation over their life cycle, can be obtained now from a contrail tracking algorithm that has been developed during QUANTIFY [34, 35]. The combination of such a tracking algorithm, based on satellite data from a polar orbiter (MODIS on NASA's Aqua and Terra satellites) and a geostationary one (MSG operated by EUMETSAT), with a numerical contrail plume model for which first developments have been made in the project allows to extend the observable fraction of a contrail life cycle in both directions, that is, from formation up to its end. There are, however, not yet trends available for a larger area with intense air traffic over Europe.

### Changes of Air Chemistry by Air Traffic

The present and future impacts of emissions from aviation are typically evaluated using several global atmospheric chemistry models in order to obtain not only a quantification of the impact of a specific transport sector on global atmospheric composition but also an estimate of the associated model uncertainties. For the QUANTIFY project such an assessment included chemistry transport models and chemistry-climate models. For these simulations the global emission inventory EDGAR32FT2000 was used for anthropogenic emissions, with the transport sectors' emissions replaced by improved estimates for

the present and for the IPCC A1 and B1 scenarios (2025, 2050) from the new QUANTIFY inventory. The A1 scenario represents a world of very rapid economic growth, especially in developing countries, and rapid introduction of new and more efficient technologies. The B1 scenario assumes rapid introduction of clean and resource-efficient technologies, and is considered to be far more optimistic with respect to environmental policy making than the A1 scenario. The results from the scenario calculations for the impact of shipping on ozone in the lower troposphere (surface to 800 hPa) point to a significantly increasing impact on ozone from the shipping sector, if emissions evolve according to the A1 scenario, while the B1 scenario leads to a much smaller increase. The results obtained for shipping are particularly interesting in the sense that the impacts on the hydroxyl radical (OH) and thus on methane (CH<sub>4</sub>) were larger than anticipated. The reduction of the global atmospheric lifetime of methane due to ship emissions was about a factor of 2–3 larger than that due to road emissions, and about a factor of 4 larger than that due to air traffic emissions [15]. This is a consequence of the more pristine chemical conditions over the oceans and the slightly different mix of emissions from ships. As a result, the net impact on radiative forcing of ozone and methane perturbations from shipping is negative, corresponding to a more or less globally uniform cooling. This adds up to the cooling via widespread cloud modifications caused by ship emissions of sulfate aerosols [27]. In contrast, the atmospheric chemistry influence on radiative forcing by aviation remains strongly positive.

The radiative forcing of aviation in the year 2000 in relation to all other transport sectors is given in Table 1 (QUANTIFY [30]). Peculiarities in comparison to other modes of transport emerge: First, the radiative forcing is nearly doubled by inclusion of the effect of ozone formation due to the emission of nitrogen oxides at cruising altitudes. Second, aviation-induced cloudiness in form of contrails and contrail cirrus adds substantially to the positive radiative forcing, however, is rather uncertain. Overall, aviation remains the sector with strongest additional effects besides the basic CO<sub>2</sub>-related radiative forcing. The line named indirect aerosol effect in Table 2 needs a special remark. It is to a large extent due to model calculations using comparably crude parameterizations for the aerosol

influence on cloud droplet number and on absorption of solar radiation by soot. This research is still in its infancy as shown by the huge error bars, and one should not argue like very often lobbies would do that any climate policy measure concerning aviation has to be postponed until small error bars are reached for all entries. What is needed is a flexible reaction should major changes become evident.

### Transport Sector's Contribution to Future Climate Change

Using a coupled atmosphere/ocean/land-model and the A1B scenario of IPCC as well as accounting for six different forcings ( $\text{CO}_2$ ,  $\text{CH}_4$ ,  $\text{O}_3$ , and aerosols for all sectors, CFCs and HCFCs for road traffic, and contrail formation for aircraft) Oliv   et al. [28] found: In the year 2100, the warming from all transport sectors due to the  $\text{CO}_2$  increase reaches about 0.5 K with the largest contribution from road traffic (0.3 K). The non- $\text{CO}_2$  impact from air traffic is larger than its  $\text{CO}_2$  impact, the latter amounting to 0.15 K at 2100. It is caused by an enhanced  $\text{O}_3$  production (compared to road and ship traffic) due to the altitude of the  $\text{NO}_x$  emissions, and by the positive direct radiative forcing from linear contrails and contrail-induced cirrus. Because the total anthropogenic global temperature increase is about 3 K in 2100 with a contribution of about 0.65 K from the transport sector, with 0.5 K due to  $\text{CO}_2$  impact and 0.15 K due to non- $\text{CO}_2$  impact dominated by aircraft traffic, the transport sector contributes slightly less than its share in total energy use. This reduced impact is caused by both the reduction in methane lifetime and comparably low  $\text{N}_2\text{O}$  and  $\text{CH}_4$  emissions. The warming is in general more pronounced at high latitudes, but the signal there is strongly influenced by the timing of the Arctic sea-ice melting, which is quite uncertain.

It is clear from the above scientific findings that international air traffic cannot be excluded from climate policy making, although it managed to remain largely unaffected so far.

### Why Did International Air Traffic So Far Escape from an Emission Reduction Policy?

When the Kyoto Protocol has been formulated in December 1997 the following sentence in its Article

2.2: “Parties included in Annex 1 (these are the industrialized countries with emission reduction commitments under the Kyoto Protocol) shall pursue limitation or reduction of emissions of greenhouse gases not controlled by the Montreal Protocol from aviation and marine bunker fuels, working through the International Civil Aviation Organization (ICAO) and the International Maritime Organization (IMO), respectively” started it. In other words: The reduction of emissions from international air traffic and international shipping has to be handled by two United Nations bodies often said to be closer to the respective industries than to the responsibility for entire society. Hence, about 6% of global emissions from the use of fossil fuels or up to 15% of all transport emissions factually escaped from any regulation by the UNFCCC and its Kyoto Protocol. Nineteen years after the signature under UNFCCC at the Earth Summit in Rio de Janeiro we are still without a global legal instrument for any mitigation measure to reduce aviation fuel’s climate impact.

Why is it so? The following lines use a historical account given by EFTE [9]. At the first Conference of the Parties (COP 1) to UNFCCC in Berlin 1995 its Subsidiary Body on Scientific and Technical Advice (SBSTA) was requested to examine the allocation and control of emissions from international bunker fuels, the term used for shipping first and then also for aviation. COP 1 had identified eight options for SBSTA to be considered prior to COP 2. SBSTA rejected three options: first, allocation in proportion to national greenhouse gas emissions; second, allocation according to the country of origin of the passenger/cargo; and third, according to the actual location of the emission. No agreement could be reached for the remaining five. There was no progress at COP 2 in Geneva. At COP 3 in Kyoto at least the domestic part of aviation and shipping was integrated into national reduction targets, and for international flights the above mentioned responsibility of ICAO was decided. Although SBSTA continued to discuss methodologies to allocate emissions, it was without any agreement so far. ICAO often reported about progress in its deliberations, but never presented a real proposal to SBSTA that could have been tabled at the next COP. Already in 1996 ICAO had requested that IPCC should prepare

a report on the impact of aircraft engines on the atmosphere. In 1999, this first report on a specific industry sector named *Aviation and the Global Atmosphere* had been published [17] that clearly stated:

- Since 1960, aviation passenger traffic had grown rapidly by 9% per year and was estimated to grow by 5% per year between 1990 and 2015.
- Already in 1992, chosen as the base year for the report, aviation accounted for 2% of total anthropogenic CO<sub>2</sub> emissions.
- Including other effects of aviation on the atmosphere, e.g., contrails and contrail cirrus, about 3.5% of the total anthropogenic radiative forcing in the base year 1992 of the report were due to aviation.

Policy options were given as well including more stringent engine regulations, removal of subsidies, environmental levies, and emission trading schemes. IPCC also acknowledged that ICAO had already started to assess the need for aircraft emission standards at cruising altitude in addition to the already existing standards for nitrogen oxides. To make a long story short: At present, 12 years later, international air traffic is not included in any binding measure for climate protection policy, i.e., ICAO was unsuccessful to perform as requested by the international community.

### The European Directive to Include Air Traffic into the Emission Trading Scheme

At least for the European Union the time for an escape of the international air traffic from an emissions trading scheme will end. The directive (Dir 2008/101/EC) to include aviation into the EU Emissions Trading Scheme (EU-ETS) was published on January 13, 2009 in the EU's Official Journal. It is valid for all flights to, from, and within the EU and starts on January 1, 2012. For 2012, an EU-wide cap on aviation emissions is set at 97% of the average annual emissions for the years 2004–2006. It will be lowered to 95% in 2013. All airlines will have to buy 15% of their allowance under the said cap through an auction. More allowances in order to cover growth of an airline have to be purchased. The revenues from selling, if emissions are below the cap, must not be put into the industry or climate protection measures. The inclusion of aviation into EU-ETS will be reviewed in 2014. There are

exemptions for very small airlines, research, military, and rescue flights. The directive has been included into the one for EU-ETS (Dir 2003/87/EC).

### An Airline Emission Index

As always in any economic sector there are large differences in behavior, here emissions, between different agents. In order to bring transparency into the aviation sector of global transport the German nongovernmental, nonprofit organization *atmosfair* has issued for the first time in spring of 2011 a global airline index, comparing nearly all airlines with respect to their environmental standards in the field of climatically relevant emissions. The *atmosfair* airline index (AAI) is a ranking of most airlines with respect to such climatically relevant emissions with the following characteristics [1]:

1. Inclusion of the 130 largest airlines in the world, arranged according to passenger kilometers as well as revenue, disregarding mere cargo flight companies.
2. Data from 2009 are used for the AAI 2011.
3. Airlines are subdivided into the following classes: net carrier, low cost carrier, charter, and regional carrier.
4. Solely climatically relevant emissions of a carrier are taken into account, disregarding noise pollution and further sustainability measures.
5. Restriction to carbon dioxide emissions, because other emissions causing for instance contrails are very similar for all airlines. There is only one exception: Nitrogen oxides (NO<sub>x</sub>) are included via an engine-related emission factor (needed to assess the radiative forcing by ozone and methane, both influenced due to chemical reactions of these oxides in the atmosphere, leading to higher ozone and lower methane concentrations).
6. Different business plans like those for net carriers and low cost carriers are not valued, despite the air traffic stimulation caused by the low cost carriers through subsidies often given to them by national governments or city councils.
7. All airlines are categorized into seven efficiency classes from A to G, in analogy to the European Union efficiency classes.



The AAI bases on a new *atmosfair* methodology, built upon the CO<sub>2</sub> emission calculations of ICAO. Its main data sources are the generally available data sets: ICAO TFS (Traffic Flight by Stage), IATA WATS, OAG, Piano-x, JP-Airline Fleet. This combination leads to 103 types of aircraft and 287 engines, constituting 87% of the global commercial air traffic with respect to number of flights. The parameters used are: aircraft type, engine, winglets, seating arrangement, cargo addition, and passenger number.

The comparison of airlines follows the following procedure:

1. Calculation of the CO<sub>2</sub> emission per kilometer for a flight between a city pair, taking into account the above parameters.
2. The CO<sub>2</sub> emission from step 1 for a city pair is compared with the best case (direct flight in best heights with least fuel consumption) and the worst case with three times higher emissions (chosen to fully embrace all airlines).
3. The airline reaching the best case for a city pair gets 100 points (it never happened), the one with higher emissions for the same city pair gets less points according to its reduced performance, reaching zero points if emitting three times as much.
4. All city pairs of an airline are combined by summing over all points reached for all pairs and dividing by the number of city pairs.
5. Categorization of all airlines into seven efficiency classes, in analogy to the color codes of EU for refrigerators, houses, etc. No single airline has reached class A or B.

This procedure allows the following comparisons:

- An airline with only a few city pairs can be compared to one servicing hundreds of city pairs.
- An airline with only short distance flights can be compared with one offering only long-distance flights.
- A low cost carrier can be compared with a net carrier.
- An airline offering only one city pair without competitors can be ranked.

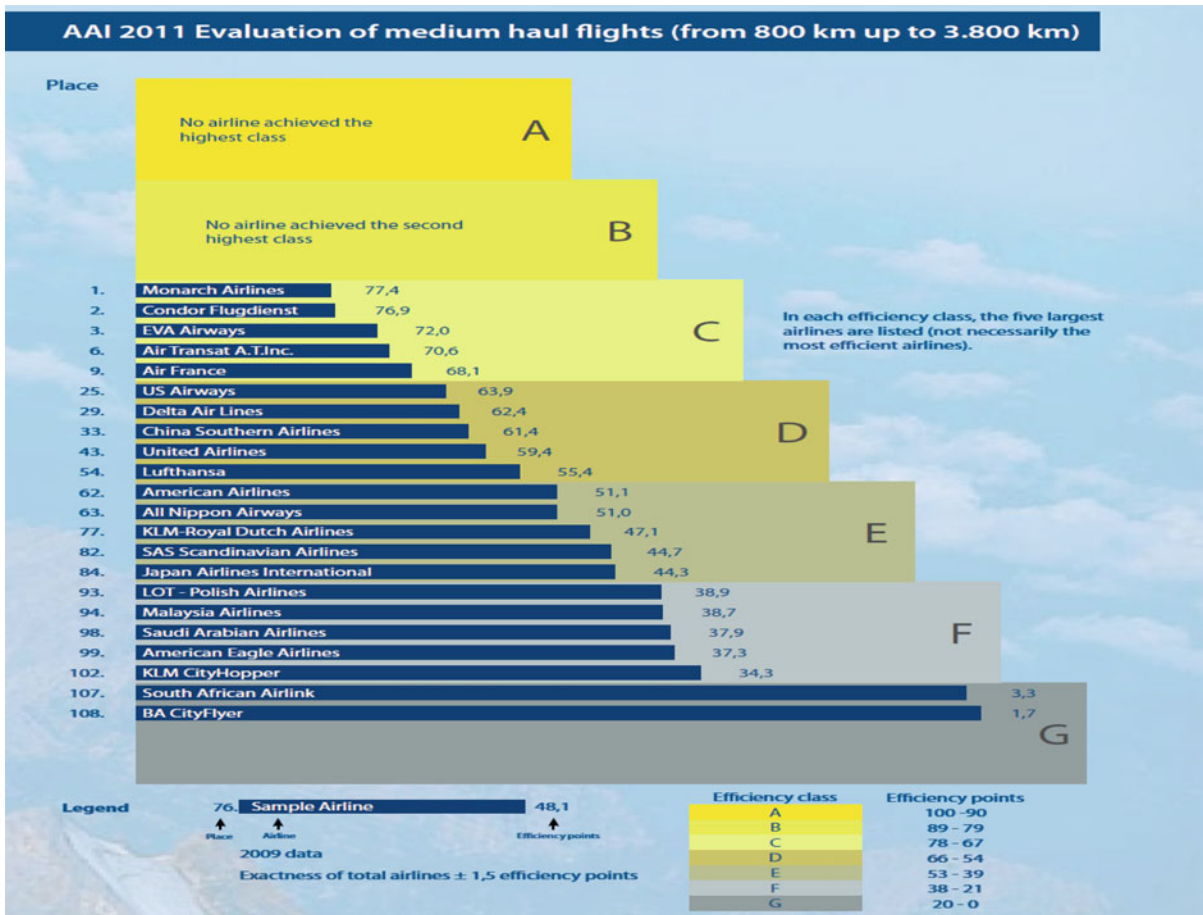
The importance of the influencing factors age of an aircraft, number of seats, and percentage of seats occupied leads to an astonishing result: Charter airlines and

low cost carriers have a lower climate impact than well-known carriers and no airline has reached category A or B. However, there needs to be a discussion on that: How should the stimulation of further air traffic by and the subsidies often given to low cost carriers be handled? The *atmosfair* group has decided not to include the low cost carriers in its first AAI although all the data are available. For the ranking of the airlines, excluding low cost carriers, see Fig. 7. There are only three factors nearly totally determining the higher ranking: high numbers of seats in a certain aircraft type, percentage of seats occupied, and age of the aircraft. Hence, an airline flying tourists to a destination with narrow seating and leased modern aircraft can reduce fuel consumption to about 3 l of kerosene per passenger and 100 km, while a net carrier with first and business class seats, some old aircraft and rather empty airplanes on certain dates for certain city pairs cannot reach higher ranks.

## Conclusions and Outlook

Mobility has a growing share in changing climate in comparison to other human activities like industrial production and heating or cooling of houses. Fortunately, the European Union's major research project named QUANTIFY was able to compare all modes of transport and very recently quantified the different modes' contributions. Taking from their conclusions and focusing here on the contribution of aviation (see QUANTIFY [30] for all modes of transport) we find:

1. The impact of transport, in particular of aviation and shipping, on climate grows faster than the impact from other sectors of human activity. Hence, aviation has to excel concerning efficiency measures in comparison to all other human activities.
2. The effects of non-CO<sub>2</sub> emissions on climate caused by aviation and shipping (e.g., through nitrogen oxide emissions and induced changes in cloudiness) are particularly large in comparison to other modes of transport.
3. A NO<sub>x</sub> molecule from aviation is responsible for five times more ozone than a molecule from road transport.
4. A NO<sub>x</sub> molecule from shipping is four times more efficient in reducing methane concentration in the atmosphere than a NO<sub>x</sub> molecule from aviation.



Aviation and Atmosphere. Figure 7

A sample from the AAI for medium distance flights. For all airlines, please see <https://www.atmosfair.de/en/air-travel-climate/airline-index/>

- Aircraft-induced clouds warm the atmosphere. Their radiative forcing (RF) is of similar magnitude than RF from aviation CO<sub>2</sub>. Hence, a multiplication factor of at least around 2 (including air chemistry effects) seems justified when dealing with climate effects of aviation.
- The relative weights of the non-CO<sub>2</sub> effects of different modes of transport strongly depend on the metric chosen and on the time horizon.
- A temperature-based climate metric leads to a very different result than one based on the radiative forcing and thus only on the greenhouse warming potential.
- The per person and per kilometer climate impact of passenger transport is much lower for railways

compared to road transport and aviation, and the per person per kilometer climate impact of passenger air travel approaches that of cars on time scale of decades. Despite all the progress achieved so far many open questions remain [30].

Most of the quantifications can only be given with partly substantial uncertainty, sometimes even effecting the sign of an effect. In particular, the climate impact of indirect aerosol effects on clouds is highly uncertain. We still do not know whether soot cirrus really exists. The net effect of NO<sub>x</sub> emission from air traffic might cause a net negative radiative forcing, if a recently proposed additional chemical reaction is confirmed. Therefore, it is presently difficult to include some non-CO<sub>2</sub> effects in climate policies related to

aviation or to develop suitable mitigation strategies for all non-CO<sub>2</sub> effects caused by the transport sector. Furthermore, many of the emitted species responsible for the non-CO<sub>2</sub> effects of air traffic emissions also impact air quality.

By applying a hierarchy of models (from process models, cloud resolving models to chemical transport models, global general atmospheric circulation models, and coupled atmosphere/ocean/land-models) and by analyzing measurements from satellites and dedicated field campaigns, the main uncertainties with respect to short-lived non-CO<sub>2</sub> effects on the atmospheric composition and on climate should be reduced:

1. The importance and magnitude of trade-offs in the NO<sub>x</sub> effect (short-term increase in ozone versus longer-term decrease in methane and the associated longer-term ozone decrease) should be explicitly simulated by comprehensive models covering the long response time of methane.
2. The climate impact of contrail cirrus should be simulated by climate models including parameterizations of contrail cirrus and evaluated by recently available tracking algorithms of these clouds in satellite data.
3. The climate impact of indirect effects on clouds (e.g., soot cirrus caused by air traffic) should be simulated by now available chemistry-climate models including aerosol-cloud interactions; the necessary parameters should be obtained by dedicated airborne measurement campaigns.
4. The regional response pattern in atmospheric composition and climate resulting from heterogeneous transport emissions need to be determined consistently in a multi-model approach.
5. The effects arising while replacing fossil fuel by biofuels should be considered. Finally, trade-offs between climate and air quality effects of emissions from different sectors should be examined for robust policy formulations.

Because the future will not only bring – with high probability – a global carbon emission trading scheme including all modes of transport but will also approach the metric “emissions per person per year” as the baseline of climate policy, the inclusion of aviation into the European Union emission trading scheme is just a first step on the way to a more just handling of the

different modes of transport. This will give a push toward a much more efficient transport system, e.g., with a preference for high speed trains in comparison to short distance flights, and it will stimulate the search for new types of fuel for aircraft derived from renewable energy sources, be it methane or hydrogen or another fuel.

## Acknowledgment

I am very thankful to Dietrich Brockhagen from *atmosfair* for many discussions over the years on attempts to integrate aviation into the emission reduction arena and especially on different metrics to account for the climate effects of air traffic besides those by CO<sub>2</sub>.

## Bibliography

1. *atmosfair* (2011) The *atmosfair* airline Index (AAI), *atmosfair* gGmbH, Berlin 2011. <https://www.atmosfair.de/en/air-travel-climate/airline-index/>
2. Balkanski Y, Myhre G, Gauss M, Rädel G, Highwood E, Shine KP (2010) Direct radiative effect of aerosols emitted by transport: from road, shipping and aviation. *Atmos Chem Phys Discuss* 10:1659–1691
3. Baughcum SL, Henderson SC, Tritz TG, Pickett DC (1996) Scheduled civil aircraft emission inventories for 1992: database development and analysis. NASA CR4700. NASA, Langley Research Center, Hampton
4. Baughcum SL, Sutkus DJ Jr, Henderson SC (1998) Year 2015 aircraft emission scenario for scheduled air traffic. NASA-CR-1998-207638. National Aeronautics and Space Administration, Langley Research Center, Hampton, 44 pp
5. Berghof R, Schmitt A, Eyers C, Haag K, Middel J, Hepting M, Gröbler A, Hancox R (2005) CONSAVE 2050. Final technical report. DLR, Köln
6. Berntsen T, Fuglestad J (2008) Global temperature responses to current emissions from the transport sectors. *Proc Natl Acad Sci (PNAS)* 105:19154–19159
7. Borken-Kleefeld J, Fuglestad J, Berntsen T (2010) Specific climate impact of passenger and freight transport. *Environ Sci Tech* 44:5700–5706
8. Cariolle D, Caro D, Paoli R, Hauglustaine D, Cuenot B, Cozic A, Paugam R (2009) Parameterization of plume chemistry into large scale atmospheric models: application to aircraft NO<sub>x</sub> emissions. *J Geophys Res* 114:D19302
9. EFTE (European Federation for Transport and Environment) (2009) Bunker fuels and the Kyoto protocol: how ICAO and the IMO failed the climate change test, Brussels, June 2009, 19 pp
10. Eyers CJ, Addleton D, Atkinson K, Broomhead MJ, Christou R, Elliff T, Falk R, Gee I, Lee DS, Marizy C, Michot S, Middel J, Newton P, Norman P, Plohr M, Raper D, Stanciou R (2005)

- AERO2K global aviation emissions inventories for 2002 and 2025. QINETIQ/04/0113, Farnborough
11. Fuglestvedt JS, Shine KP, Berntsen T, Cook J, Lee DS, Stenke A, Skeie RB, Velders GJM, Waitz IA (2010) Transport impacts on atmosphere and climate: metrics. *Atmos Environ*. doi:10.1016/j.atmosenv.2009.04.044
  12. Gardner RM, Adams JK, Cook T, Larson LG, Falk RS, Fleuit E, Förtsch W, Lecht M, Lee DS, Leech MV, Lister DH, Massé B, Morris K, Newton PJ, Owen A, Parker E, Schmitt A, ten Have H, Vandenberghe C (1998) ANCAT/ EC2 aircraft emissions inventories for 1991/1992 and 2015. Final report. Produced by the ECAC/ANCAT and EC working group. European civil aviation conference
  13. Henderson SC, Wickrama UK, Baughcum SL, Begin JJ, Franco F, Greene DL, Lee DS, McLaren ML, Mortlock AK, Newton PJ, Schmitt A, Sutkus DJ, Vedantham A, Wuebbles DJ (1999) Aircraft emissions: current inventories and future scenarios. In: Penner JE, Lister DH, Griggs DJ, Dokken DJ, McFarland M (eds) 'Aviation and the Global Atmosphere', intergovernmental panel on climate change. Cambridge University Press, UK
  14. Hodnebrog Ø, Berntsen TK, Dessens O, Gauss M, Grewe V, Isaksen ISA, Koffi B, Myhre G, Olivie D, Prather M, Stordal F, Szopa S, Tang Q, van Velthoven P, Williams J, Ødemark K (2011) Future impact of nonland-based traffic emissions on atmospheric ozone and OH – an optimistic scenario and a possible mitigation strategy. *Atmos Chem Phys Discuss* 11:16801–16859
  15. Hoor P, Borken-Kleefeld J, Caro D, Dessens O, Endresen O, Gauss M, Grewe V, Hauglustaine D, Isaksen ISA, Jöckel P, Lelieveld J, Myhre G, Meijer E, Olivie D, Prather M, Schnadt Poberaj C, Shine KP, Staehelin J, Tang Q, van Aardenne J, van Velthoven P, Sausen R (2009) The impact of traffic emissions on atmospheric ozone and OH: results from QUANTIFY. *Atmos Chem Phys* 9:3113–3136
  16. IEA (2007) Oil Information 2006, Table 9, 749 pp. International Energy Agency, Paris
  17. IPCC (1999) Aviation and the global atmosphere. In: Penner JE, Lister DH, Griggs DJ, Dokken DJ, McFarland M (eds) Intergovernmental panel on climate change. Cambridge University Press, Cambridge
  18. IPCC (2000) In: Nakicenovic N, Swart R (eds) Emission scenarios. Cambridge University Press, Cambridge, pp 570
  19. IPCC (2001) Climate change 2001: the scientific basis. Report of working group I. Cambridge University Press, Cambridge, UK
  20. IPCC (2007a) In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change, 2007. Cambridge University Press, Cambridge/New York
  21. IPCC (2007b) In: Parry ML, Canziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change, 2007. Cambridge University Press, Cambridge/New York
  22. Kim BY, Fleming GG, Lee JJ, Waitz IA, Clarke JP, Balasubramanian S, Malwitz A, Klima K, Locke M, Holsclaw CA, Maurice LQ, Gupta ML (2007) System for assessing Aviation's Global Emissions (SAGE), Part 1: model description and inventory results. *Trans Res D* 12:325–346
  23. Koffi B, Szopa S, Gozic A, Hauglustaine D, van Velthoven P (2010) Present and future impact of aircraft, road traffic and shipping emissions on global tropospheric ozone. *Atmos Chem Phys* 10:11681–11705
  24. Lee DS et al (2009) Aviation and global climate change in the 21st century. *Atmos Environ*. doi:10.1016/j.atmosenv.2009.04.024
  25. Lee DS et al (2010) Transport impacts on atmosphere and climate: aviation. *Atmos Environ* 44:4678–4734
  26. Liu X, Penner JE, Wang M (2009) Influence of anthropogenic sulphate and black carbon on upper tropospheric clouds in the NCAR CAM3 model coupled to the IMPACT global aerosol model. *J Geophys Res* 114:D03204. doi:10.1029/2008JD010492
  27. Myhre G, Shine KP, Rädel G, Gauss M, Isaksen ISA, Tang Q, Prather M, Williams JE, van Velthoven P, Dessens O, Koffi B, Szopa S, Hoor P, Grewe V, Borken-Kleefeld J (2010) Radiative forcing due to changes in ozone and methane caused by the transport sector. *Atmos Environ* 45:387–394
  28. Olivie DJL, Cariolle D, Teyssède H, Salas D, Voldoire A, Clark H, Saint-Martin D, Michou M, Karcher F, Balkanski Y, Gauss M, Dessens O, Koffi B, Shine KP, Sausen R (2010) Modeling the climate impact of road transport, maritime shipping and aviation over the period 1860–2100 with an AOGCM. Submitted to *Atmos Chem Phys*
  29. Penner JE, Chen Y, Wang M, Liu X (2009) Possible influence of anthropogenic aerosols on cirrus clouds and anthropogenic forcing. *Atmos Chem Phys* 9(2009):879–896
  30. QUANTIFY (2011) Quantifying the climate impact of global and European transport systems, integrated project, sixth framework programme, sub-priority 1.1.6.3: global change and ecosystems, final activity report, publishable summary
  31. Sausen R (ed) (2010) Transport impacts on atmosphere and climate: the ATTICA assessment report. *Atmos Env* 44, 4645–4816 (special issue)
  32. Sausen R, Schumann U (2000) Estimates of the climate response to aircraft CO<sub>2</sub> and NO<sub>x</sub> emissions scenarios. *Clim Change* 44:27–58
  33. Sutkus DJ, Baughcum SL, DuBois DP (2001) Scheduled civil aircraft emission inventories for 1999: database development and analysis. National aeronautics and space administration, glenn research centre. NASA CR-2001/211216
  34. Vazquez-Navarro M (2010) Life cycle of contrails from a time series of geostationary satellite images. *DLR-Forschungsbericht* 2010–19, 139 pp
  35. Vazquez-Navarro M, Mannstein H, Mayer B (2010) An automatic contrail tracking algorithm; *Atmos. Meas Tech* 3:1089–1101

